LEGUME PHYLOGENOMICS AND THE LIMITS TO PHYLOGENETIC RESOLUTION

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Erik J.M. Koenen

aus

die Niederlande

Promotionskommission

Prof. Dr. H. Peter Linder (Vorsitz)
Dr. Colin E. Hughes (Leitung der Dissertation)
Prof. Dr. Kentaro Shimizu
Prof. Dr. Michael J. Sanderson (externes Gutachten)
Dr. Pascal-Antoine Christin (externes Gutachten)

Zürich, 2019

Summary

The field of phylogenomics involves estimating species relationships using genome-scale data and investigating genome evolution, including processes such as incomplete lineage sorting (ILS), introgression through hybridization, and (paleo)polyploidy or whole genome duplication (WGD). Phylogenomic studies are revealing the full complexity in genome evolution caused by these processes which can result in many different evolutionary histories for genes and other genomic elements contained within each genome. These conflicting evolutionary gene histories across genomes suggest that organismal evolution does not always follow a strictly bifurcating tree-like pattern and that reticulation and hard polytomies are relatively common.

In this thesis, I apply phylogenomic analyses to newly generated genome-scale data sets for the angiosperm family Leguminosae (or Fabaceae), to enhance our estimate of legume phylogeny and to better understand what processes underly the difficulties in resolving parts of the legume phylogeny. The legumes are the third largest family of angiosperms with c. 19,581 described species in 770 currently recognised genera. It is a clade of tremendous ecological and economic importance, being widespread and often dominant across the globe (except in Antarctica) in various tropical and temperate forest and grassland ecosystems and it is the second most cultivated plant family after the grasses (Poaceae). Despite this importance, several parts of the legume phylogeny remain poorly characterized, including the deepest divergences in the family (i.e. the relationships among the six subfamilies) and a large portion of the mimosoid clade of subfamily Caesalpinioideae. The mimosoid legumes, with c. 3,300 species, form a prominent clade of lowland pantropical woody legumes that occur abundantly and often dominantly in rainforests, savannas, seasonally dry tropical forests and semi-arid scrublands. Within the mimosoids, the Ingioid clade forms the largest poorly-resolved part of the legume phylogeny and its (supra-)generic classification has long been problematic with generic delimitation in a state of considerable flux.

In Chapter I, I demonstrate that the six major lineages of legumes that are currently recognized as subfamilies originated nearly simultaneously. Data sets used consist of a matrix of 72 protein-coding chloroplast genes for 157 taxa and a set of alignments of 9,282

gene clusters for 76 taxa. Concatenated phylogenetic analyses, gene tree summarization methods and supernetwork reconstruction are applied showing that there is a lack of phylogenetic signal across the majority of gene trees for the earliest divergences in the legume phylogeny, and revealing strongly conflicting relationships in the remainder of gene trees, suggestive of ILS or introgression. This implies that the deepest divergences in the family occurred in rapid succession. This near-simultaneous origin of the six legume subfamilies has important implications for understanding the evolution of legume diversity and traits. The prevailing view that some subfamilies are “basal” or “early-diverging” with respect to others should be abandoned, with each subfamily being equally likely to have maintained ancestral (plesiomorphic) traits or to have evolved derived (and potentially homoplasious) traits.

Using the nuclear gene tree set from Chapter I, in Chapter II I investigate the occurrence of polyploidy early in the evolution of the Leguminosae and the placements and timing of multiple putative WGDs, as well as the timing of the initial radiation of the legumes relative to the Cretaceous-Paleogene (K-Pg) boundary (KPB). Using gene tree reconciliation methods, gene count data and supernetwork reconstruction, I tested the number of WGDs that likely occurred and their placements among the earliest divergences in the family, as well as whether there is evidence for allopolyploidy. While some gene tree reconciliation results suggest a pan-legume WGD prior to the first divergences in the family in addition to WGDs specific to subfamilies Detarioideae and Papilionoideae, other analyses suggest an allopolyploid origin of (at least) subfamily Caesalpinioideae that is potentially shared with Papilionoideae, suggesting that the latter subfamily underwent two rounds of legume WGD prior to its initial diversification. The allopolyploid scenario is considered more likely since gene tree reconciliation methods that do not account for allopolyploidy can be misled in inferring an earlier WGD at the divergence of the two parental lineages of the allopolyploid. Time-calibration analyses show that both the initial diversification of the legumes and the WGDs, with the exception of the Detarioideae WGD, are closely associated with the KPB. Taken together, these results suggest that both ancient polyploidy and evolutionary turnover at the KPB due to mass extinction played a role in the rapid initial diversification of the family and its abundance in (early) Cenozoic fossil assemblages, by providing expanded genomic

substrates for diversification and ecological opportunities in newly emerging post-KPB habitats in the Paleocene.

Chapter III describes a complete hybrid capture project aimed at enhancing our estimate of the phylogeny of the currently poorly resolved mimosoid clade of subfamily Caesalpinioideae. The methods used here involved: (1) generating genomic data for four mimosoid genera with RNAseq; (2) selecting putative low-copy nuclear genes to target; (3) preparing sequencing libraries that are enriched for the targets using hybrid capture; and (4) data assembly and phylogenomic analysis. The first and third parts use relatively well-known methods that are well-described in the scientific literature, while the second and fourth parts were specifically designed for this study and implemented using a combination of custom python scripts and existing bioinformatic and phylogenetic tools, which are described in detail. The final data set for phylogenomic analyses consists of 1,767 gene alignments, and includes sequence data for 122 taxa. Results show that this hybrid capture protocol can successfully generate genome-scale data across the mimosoid clade, and results in a much improved estimate of the mimosoid phylogeny. I demonstrate that several mimosoid genera, including the ‘dustbin’ genus *Albizia,* and most informal groupings within the Ingioid clade, are not monophyletic. These results form the basis for future taxonomic work, including a new tribal and clade-based classification for subfamily Caesalpinioideae in which mimosoids are included. Several cases of gene tree conflict across the phylogeny are further analysed, most notably showing that within the large pantropical Ingioid clade of c.38 genera and more than 2,000 species, there is a near-complete lack of phylogenetic signal across gene trees involving relationships among 6 or 7 well-supported Ingioid subclades. This suggests that there is a hard polytomy embedded within the initial radiation of the Ingioid clade involving 6 or 7 lineages. Using the same hybrid capture protocol to increase taxon sampling will lead to a large and well-resolved phylogeny for Caesalpinioideae in the near future, providing the basis for future studies on macro-evolution and biogeography of this subfamily.

This thesis forms a significant contribution to legume phylogenomics and to understanding processes that lead to poor phylogenetic resolution and non-bifurcating organismal evolution more generally. Given the ecological importance of legumes, their initial diversification following the KPB and their abundance throughout the Cenozoic, the findings in

this thesis contribute important evidence for explaining the origins of Cenozoic angiosperm diversity. Genome evolution in legumes is shown to be highly complex, and by explicitly addressing the roles of rapid lineage diversification, ILS, paleopolyploidy, hybridization and methodological issues causing gene tree discordance, this thesis highlights the limits to phylogenetic resolution.

Preface

My thesis is finally finished. I am extremely happy and thankful for that, after all the hard work and the ups and downs, to have accomplished writing the three manuscripts presented in Chapters I-III, and to have completed my PhD thesis, was really worth the effort in the end. I also think I have been very lucky to have had the opportunity to do a PhD on such an interesting topic and especially to have met so many nice and interesting people along the way. My PhD has taken me to several far corners of the world, for conferences, field work and collaborations. Field work in Brazil and Madagascar have been amazing experiences, especially seeing such disparate tropical ecosystems as NE Brazilian Caatinga and the seasonally flooded forests of the Upper Rio Negro, or the Malagasy semi-arid thorn-scrub and montane forests of Marojejy, has really been very inspiring for my research on the evolution of legume diversity. I have also really enjoyed collaborating with researchers in Feira de Santana, New York, Brussels, Wageningen, Leiden, London and above all with the *Inga* research team in Edinburgh. This collaboration with Toby, Catherine, James and Kyle has been started shortly after starting my PhD and has been tremendously important for broadening my knowledge and skills and it has been central to the work presented in Chapter III. I have always very much enjoyed visiting and working in Edinburgh at the Botanics. Another part I really enjoyed was visiting and working in several important large herbaria around the world, studying their amazing collections. I also had the opportunity to supervise a Masters student, Anahita Aebli, which was a very nice experience from which I learned a lot. Anahita was a very talented student, who also came to Madagascar for part of the fieldwork that I did there and she produced an excellent thesis. Apart from being lucky to have had all these experiences during my PhD, I think I am also lucky for the study group that I happened to have been working on for this thesis, the Leguminosae, an exceptionally interesting plant family with a wonderful collaborative international research community devoted to systematic research on it.

The final composition of this thesis in terms of subject area and the scientific questions that are being tackled is rather different from what was envisioned at the start of the project. The project was originally focused on the mimosoid legumes, of which the third chapter is still testimony, while the first two chapters take the whole of the legume family as study group. The demise of subfamily Mimosoideae is not the reason for this, as a change in taxonomic

status for the clade does not mean it is no longer worth studying them of course. The two main reasons, I believe, for making this shift during the project are related to technical issues and because of following an interesting side alley further than I had planned to.

The technical issues arose when I was trying to do hybrid capture, a pioneering new technique at the time but by now steadily becoming the method of choice for phylogenomic studies on non-model taxa. Several failures in the lab made me feel rather demotivated and I turned to analysing some chloroplast genome and transcriptome data that I had partly generated myself and tried inferring the basal relationships in the legume family which was a difficult phylogenetic problem for the legume research community for some time already. When later the hybrid capture work did yield some data after I moved to work in the lab at the Functional Genomics Center at the Irchel campus in Zurich, it was foreseen that the first chapter of my thesis would be on basal relationships in the legumes and the other two chapters would focus on mimosoid phylogenomics and macroevolution. After three years in Zurich, I left to go to the Netherlands for a while, as well as spending some time at the Royal Botanic Gardens in Edinburgh. During this time, I continued working on both the basal legume relationships and the mimosoid data.

Eventually, the manuscript on deep divergences in legumes, after including analyses on ancient polyploidy and the associations with the Cretaceous-Paleogene boundary, grew to a behemoth which was not particularly well-suited for publication in a scientific journal, which finally dawned on us after two submissions and one round of peer review. As the reviewers and editor suggested, we decided to split the manuscript in two. One of the two resulting manuscripts is included here as Chapter I and has been submitted to the journal New Phytologist at the time of handing in this thesis. The other manuscript has been amended to include more in-depth analyses of ancient polyploidy while keeping, for the moment, most of the rather lengthy introduction and discussion in the version that is presented here as Chapter II. It will most certainly be modified before publication (foreseen to be published in Systematic Biology) after co-authors and reviewers have commented on it, and will most likely be edited as well to be more concise, moving some sections to Supplementary Information. The manuscript presented in Chapter III is also foreseen to be published with some modifications, including length reduction.

Apart from the main three chapters of this thesis, I have contributed to several other studies during the course of my PhD. Most significantly, I was part of the core group of authors of the article by the Legume Phylogeny Working Group (LPWG) on the reclassification of the legume subfamilies, published in 2017. In this article, we included the most comprehensive legume phylogeny to date based on the chloroplast locus *matK*, for which I curated the mimosoid sequence data, made the final sequence alignment for all legume accessions, and inferred the Maximum Likelihood and Bayesian estimates of the phylogeny. Furthermore, I wrote the formal description of the re-circumscribed subfamily Caesalpinioideae which now includes the mimosoids, and contributed to the writing of other sections of the article. Two other articles that I co-authored with LPWG were a review of legume phylogenetics and classification and a progress report on the process of deciding how to delimit the legume subfamilies in consultation with the legume systematics research community.

Other studies that I contributed to include three articles on hybrid capture. One of these articles, on the genus *Inga* (Nicholls et al., 2015), came out of the collaboration with Toby Pennington's research group in Edinburgh and the third chapter of this thesis forms the counterpart of that article with a focus on the whole mimosoid clade. For the other two hybrid capture studies, I designed the set of targeted loci using the same procedure as described in Chapter III (Couvreur et al., 2019; Ojeda et al., 2019).

One further paper focuses on chloroplast genomes of mimosoid legumes, for which I assembled the *Inga* chloroplast genome from which I first discovered the expansion of the inverted repeat (IR) in Ingioid plastomes (Dugas et al., 2015). And finally, early on in my PhD I contributed to a study on legume macroevolution, sharing first authorship with a former colleague from Zurich, Jurriaan de Vos (Koenen & de Vos et al., 2015). Abstracts of all articles to which I contributed during my PhD are included in Appendix I.

Acknowledgements

I would like to use this preface to express my sincere gratitude to the many people that supported and helped me during my PhD studies. First and foremost, I want to thank my supervisor Colin Hughes who has tirelessly supported and motivated me throughout all these

years. I really could not have wished for a better supervisor and I am very grateful for the inspiring discussions about legume systematics and macroevolution. I would also very much like to thank my colleagues from throughout the years, especially Guy Atchison, Renske Onstein, Yanis Bouchenak-Khelladi, Yaowu Xing and Tommy Nymann with whom I shared an office for most of the time I spent in Zurich and who became good friends. It was wonderful getting to know you, talking about science (ideally over a beer) and making excursions to the Alps and the Langstrasse. I also want to thank Anahita Aebli for working together on Malagasy mimosoids. And I wish to thank Jens Ringelberg for working together on continuing the mimosoid phylogenetics work, I look forward to work on this over the next few years as well. I would like to thank Peter Linder for the inspiring scientific discussions, journal clubs, being on my PhD committee and the opportunity to assist with teaching the Swiss flora course, which I very much enjoyed. Florian Schiestl and Elena Conti are thanked for taking care of our institute as directors during the course of my PhD, and the former as well for additional funding to extend my PhD. Corinne Burlet, Claudio Brun deserve gratitude from many people, I would like to thank them as well for the essential administrative work. I also would like to thank Martin Spinnler for his great work in the library, always willing to do his best to find the literature I was looking for. I also wish to thank Barbara Keller very much for making sure things were running smoothly in the lab and all the assistance she gave for my molecular lab work. Reto Nyffeler and the other herbarium staff are thanked for handling my loans. And the garden staff, in particular René Stalder, Markus Meierhofer and Manfred Knabe for taking care of the living plants that I used during my research. In general I would like to thank the whole institute of Systematic Botany (or Department of Systematic and Evolutionary Botany as it is now called), all the sientific, technical, cafetaria and garden staff and students that I have interacted with over the years and for making the institute and gardens such a nice place in Zurich.

I would also like to thank the Functional Genomics Center (FGCZ) and the Genetic Diversity Center (GDC) in Zurich for access to their user labs, and the S3IT for the ScienceCloud computational infrastructure which was very much essential for being able to carry out this research. Of the FGCZ, I especially would like to thank Catherina Aquino who helped me in the lab and without whom I never would have been able to generate the data presented in Chapter III.

Apart from in Zurich, there are also people all over the world that have been in some way important for me or my thesis over the last years. Many of these people are part of the global legume systematics research community, most of them members of the Legume Phylogeny Working Group as well. I would first very much like to thank Toby Pennington, Catherine Kidner, James Nicholls and Kyle Dexter, from the Royal Botanic Gardens Edinburgh and the University of Edinburgh, for the amazing collaboration and all the nice times I spent there, and I wish to thank Flavia Pezzini for hosting me in Edinburgh.

I also wish to thank the Brazilian legume research community, and especially Luciano Paganucci de Queiroz, Elvia Souza, Petala Ribeiro, Marcelo Simon, Joao Iganci, Marli Morim, Francis Bonadeu and Haroldo Lima for great times spent doing fieldwork around Brazil, for ongoing collaboration on mimosoid systematics and the nice evenings and nights spent eating mocequas, drinking caipirinhas and dancing to Samba and Forró.

Melissa Luckow is thanked for sending a whole load of mimosoid leaf samples and kindly receiving me as a guest at Cornell University and hosting me in Ithaca, New York. Her work and our collaboration have been extremely important for my research on mimosoids, as presented in Chapter III and in ongoing projects that I am working on. I also wish to very much thank Gwilym Lewis, for being one of the most inspiring legume researchers out there and especially for making me feel welcome at Kew multiple times, helping me out while there and advising me on several topics such as mimosoid taxonomy, career paths and life in general. And I want to send many thanks to Anne Bruneau, for the very enjoyable and important collaborations on legume classification, phylogenomics and ongoing work on hybrid capture in Caesalpinioideae.

Jan Wieringa and Freek Bakker are thanked for giving me the opportunity to visit Wageningen University and the herbaria of Wageningen and Leiden several times and for collaboration on legume phylogenomics.

I would also like to thank Luciano de Queiroz and Domingos Cardoso for inviting me for a legume symposium at the Latin American Botanical Congress in Salvador da Bahia, Brazil, Kyle Dexter for inviting me for a symposium at the annual meeting of the Association of Tropical Biology and Conservation in Montpellier, France, Tingshuang Yi for inviting me for a symposium at the International Botanical Congress in Shenzhen, China and especially

Tadashi Kajita for inviting me to give a plenary talk at the 7th International Legume Conference in Sendai, Japan.

I also wish to thank Yaowu Xing and his new colleagues at the Xishuangbanna Tropical Botanical Garden for an amazing research visit to his new lab.

The herbarium staff of Kew, Edinburgh, Geneva, Wageningen, Leiden, Paris, Florence, New York, Rio de Janeiro, Brasilia, Pretoria and Kunming are thanked for their assistance while visiting these herbaria.

And I would very much like to thank my family and friends in the Netherlands for supporting me all these years: my father Jos, my mother Ans and my brother Mark; Raoul, Harmen, Melle, Dimitri, Mathieu, Mara, Anaïs, Martine, Lindy and all the other nice and crazy people from Droevendaal and beyond!

I am also very grateful to my girlfriend, Caroline, who I met at a conference where we were both presenting some of the work of our PhDs, so it was in fact our PhDs that made us meet each other. Being with her transformed my life and she was very important for not letting me give up, I am not sure I would have been able to finish this thesis without her. Thank you so much for all the love and kindness you gave to me and the wonderful moments we shared during travelling, while living together in Paris or when we were in Rio de Janeiro and Wageningen!

And finally I would like to thank Switzerland, including its National Science Foundation and the Claraz Shenkung, for giving me the opportunity to pursue a scientific career through funding and the magnificent Swiss Alps in which I enjoyed botanising, hiking, mountainbiking and snowboarding which greatly increased my well-being in between all the hard work on this thesis.

Table of contents

The very nature of the evolutionary history of organisms and the limitations of current phylogenetic reconstruction methods mean that part of the tree of life might prove difficult, if not impossible, to resolve with confidence. **- Delsuc et al., 2005.**

Students of animals and plants have long accepted that incomplete lineage sorting, introgression, and full-species hybridization pose difficulties for the sorts of trees that Darwin might have had us draw. **- Doolittle & Brunet, 2016.**

Introduction

The Earth's biodiversity is vast, with astronomically high estimates of the total number of species ranging from 3 – 100 million. Perhaps the total number of eukaryotes is around 8.7 million (Mora et al., 2011), of which some 86% remain undescribed. So diverse is the living world that for the human mind it is hard to grasp its true extent, and probably few people are aware of the true scale of species diversity on the planet. To give some examples of estimates of species diversity in particular taxonomic groups or ecological communities can help to more clearly envisage with just how many other distinct types of organisms we share the Earth. For instance, it has been estimated that a single gram of forest soil may contain between 12,000 and 18,000 distinct genomes (strains or "species") of bacteria (Torsvik et al., 1996). Above ground, a single hectare of Amazonian rainforest may contain 473 tree species (Valencia et al., 1994), while the total known seed plant diversity of the Amazon basin in its entirety is estimated at 14,003 species, of which 6,727 are trees (Cardoso et al., 2017). The Angiospermae are by far the largest group of plants, and while not as diverse as Arthropoda, the diversity of which runs into the millions, Angiospermae is an enormously diverse clade with estimates of the total number of species ranging from c. 260.000 (Thorne, 2002; Scotland & Wortley, 2003) to c. 369.000 (Kew, 2016). Within the angiosperms, the Leguminosae, which are the focus of this thesis, form the third largest family with nearly 20,000 published and accepted species (LPWG, 2017). It is also the family that contains the largest genus of plants, *Astragalus*, commonly known as the milk-vetches, with an estimated 2,500-3,000 species mainly distributed across the Northern Hemisphere. Because of this

spectacular diversity on the planet, cataloguing the Earth's living organisms has been and remains both a major goal of, and one of the biggest challenges in biology ever since the first naturalists started collecting specimens and emergence of the research field of taxonomy.

A related central goal of biology has been to estimate how all of this diversity of life evolved and how taxa are related to each other, generating the metaphor of the *Tree of Life* (ToL) – a gigantic bifurcating tree-like diagram composed of nodes and edges (or internodes/branches) that connect all of the millions of extant species with each other and with their extinct relatives and ultimately, with the last universal common ancestor. Drawing trees to represent a classification of organisms goes back to pre-Darwinian times, but most of these are not true phylogenies as they do not indicate relationships based on common descent. After Darwin's publication of the Origin of Species (1859), the use of phylogenies to illustrate the evolution of organisms became more common. With the introduction of cladistics (Hennig, 1966), phylogeny reconstruction became a methodology to infer natural classifications of organisms. Then, with the advent of DNA sequencing, phylogenetic systematics truly took off, leading to large-scale changes in the classification of organisms (e.g. APG, 1998, 2003, 2009; APG et al., 2016) and a flourishing of methodological developments for analysis of molecular sequence data to infer phylogenies (see Yang & Rannala (2012) for a review). However, the use of genomic data also led to problems with the ToL metaphor, as it was discovered that due to extensive lateral gene transfer (LGT), microbial evolution is poorly represented by a bifurcating tree and that a network or Web of Life is a more accurate representation of relationships among Bacteria and Archea (Doolittle, 1999; Doolittle & Brunet, 2016). Moreover, eukaryote evolution is not always well represented by a strictly bifurcating tree, since in eukaryotes there is also widespread (but less extensive) reticulation caused by hybridization (e.g. Marcet-Houben & Gabaldón, 2015; Meier et al., 2017) and LGT (e.g. Christin et al., 2012; Li et al., 2014; Dunning et al., 2019). Furthermore, deep coalescence leading to incomplete lineage sorting (ILS) appears to be common in eukaryote evolution, especially associated with episodes of rapid successive speciation (e.g. Suh et al., 2015), and also leads to gene tree discordance with respect to the species tree. The realisation that gene trees do not necessarily (or most often do not) resemble the species tree, as well as the abandonment of the ToL hypothesis in microbiology, has led to

suggestions that the evolution of life on earth is best considered as a forest of gene trees (Koonin & Wolf, 2009). However, "*in multicellular eukaryotes the molecular mechanisms and species-level population genetics of variation do indeed mainly cause a tree-like structure over time*" (Bapteste et al., 2009). In other words, the eukaryotic ToL can be seen as an emergent organismal phylogenetic tree that is deeply rooted in an organismal web of Bacteria and Archea (Doolittle, 1999), that connects the forest of microbial gene trees. And while the phylogeny of multicellular Eukarya is far from being free of reticulation and is perhaps rife with hard polytomies, it is for the largest part still well represented by a bifurcating organismal species tree, within which gene trees reflect more complex incongruent patterns of gene evolvution, even if not a single individual gene tree is fully congruent with the organismal tree. Perhaps in particular for macroevolutionary studies in eukaryotes, the ToL is a useful metaphor and research tool for gaining insights into the processes responsible for the evolution of the enormous eukaryotic species diversity on the planet, regardless of the amounts of deep-coalescence and reticulation among closely related lineages.

This thesis presents several new and important contributions to the catalogue of Life on Earth, providing a new phylogenetic framework for legume evolution (Chapter I of this thesis), lending further support to the reclassification of the legume subfamilies by the Legume Phylogeny Working Group (2017), as well as presenting a new protocol and initial phylogeny for studies on mimosoid legumes (Chapter III of this thesis), which will lead to large-scale re-delimitation of mimosoid genera and a new classification for Caesalpinioideae in the near future. By tackling some of the most difficult phylogenetic problems in the prominent legume family, including the deepest dichotomies and the least resolved clade in the family, this thesis also significantly contributes to the goal of reconstructing the ToL.

The recent merging of the fields of phylogenetics and comparative genomics – referred to as phylogenomics – has led to a much broader and fuller understanding of the origins of biodiversity, by comparing genome-scale data for tens to hundreds of genomes and reconstructing their evolution in a phylogenetic framework. In particular, phylogenomics is starting to reveal the true extent to which processes such as ILS, LGT, hybridization and paleopolyploidy (or whole genome duplication, WGD) have shaped genome evolution across the eukaryote ToL. In early legume evolution, multiple polyploid events occurred and Chapter

II of this thesis presents significant new insights into the role of polyploidy in legume evolution and the difficulties of obtaining phylogenetic resolution along the legume backbone. More generally, Chapter II presents important insights into how polyploidy and massive biotic turnover at the K-Pg boundary shaped the evolution of angiosperm diversity, separately, as well as in concert. In the next paragraphs, I first review recent developments in phylogenomics, then introduce the study groups of this thesis, namely the legume family and the mimosoid clade, and finally I outline the main chapters and hypotheses that are tested within these.

Recent developments in phylogenomics

The next-generation sequencing (NGS) revolution has led to a rapid increase in complete genome sequencing, alongside the introduction of targeted massively parallel sequencing techniques such as RNAseq, hybrid capture and RADseq, meaning that the volumes of available sequence data are exploding. The term ‘phylogenomics’ was coined in the late 1990s (Eisen, 1998; O’Brien & Stanyon, 1999). According to Philippe et al. (2005), “*the main issues are (a) using molecular data to infer species’ relationships and (b) using information on species’ evolutionary history to gain insights into the mechanisms of molecular evolution.*” I give some examples of recent studies and methodological developments relating to these two issues.

Some of the first phylogenetic studies to employ genome-scale data were on yeast species, where with a matrix of 106 genes, Rokas et al. (2003) inferred a highly supported species tree despite widespread incongruence across gene trees. Yeasts have continued as a model system for phylogenomics, and a more recent data set which includes 1,070 gene alignments that yield 1,070 distinct gene tree topologies that are all different from the species tree (Salichos & Rokas, 2013), highlighted gene tree incongruence even more strikingly. With the advent of genome-scale data it wasinitially thought that using larger sequence data sets would solve incongruence problems by reducing random and sampling errors, but conflicting results using genome-scale data, for example on the early diversification of animals (Philippe et al., 2011), showed that this would not necessarily be the case. Indeed, phylogenomics has

not solved, but instead highlighted the complexity of many difficult phylogenetic problems, such as the controversies surrounding the relationships within placental mammals (Teeling & Hedges, 2013); the Neoaves clade of birds (Suh, 2016) and in the land plant phylogeny, where the mono- or paraphyly of bryophytes is still heavily debated (Qiu et al., 2006; Wickett et al., 2014; de Sousa et al., 2019).

Given the sparse availability of complete genome sequences, many early phylogenomic studies focused on deep relationships in clades for which sequencing effort had been relatively high, but the introduction of targeted sequencing in NGS has opened up opportunities to conduct phylogenomic studies on non-model groups. Several reviews of targeted sequencing and the different techniques used in many different studies, have been written (e.g. Cronn et al., 2012; Lemmon & Lemmon 2013), of which I mention a few here. RNAseq, although not specifically a targeted sequencing technique can be used in a similar way to create ‘reduced representation sequencing libraries’, meaning that particular regions of the genome, in this case expressed genes in a particular tissue, are enriched in the sequencing library prior to sequencing. And indeed, RNAseq has been used to generate data for phylogenomics in several plant groups (e.g. Caryophyllales, Yang et al. 2015; land plants, Wickett et al., 2014; and legumes, Cannon et al., 2015 and Chapters I & II of this thesis). A different approach is offered by the use of restriction enzymes in methods such as RADseq and GBS, where regions adjacent to the restriction site are sequenced (e.g. Eaton & Ree, 2013; Wagner et al., 2013; Cavender-Bares et al., 2015; Atchison et al., 2016). However, perhaps the most widely used current method is hybrid capture, where single-stranded bait molecules (either DNA or RNA) are hybridized in solution to the complementary targeted regions and subsequently captured, usually with magnetic beads, while the unwanted regions are (for the most part) washed away (Grover et al., 2012). Usually the targeted regions are the exons of low-copy and/or functionally interesting genes (e.g. Mandel et al., 2014; Nicholls et al., 2015; Sass et al., 2016; Moore et al., 2017; Couvreur et al., 2018; Johnson et al., 2018; Chapter III of this thesis), but regions conserved across taxa regardless of function have also been targeted specifically for phylogenomics (e.g. ultra-conserved elements (UCE) loci, McCormack et al., 2012; “anchored phylogenomics”, Lemmon et al., 2012). An added benefit of hybrid capture is that organellar or ribosomal DNA, which exist in high copy numbers, can

often be extracted and analysed from off-target sequencing reads (Weitemier et al., 2014). In Chapter III, a complete hybrid capture project, from generating genomic data to target selection and analysis of the captured sequence data, is described for the mimosoid legumes, which includes the most difficult to resolve large clade within the legume family, the Ingioid clade.

In terms of analysis methods, particularly important developments have taken place to analyse and/or accommodate gene tree incongruence. These include Bayesian species tree analyses with a full multi-species coalescent model within which gene trees evolve and do not necessarily match the simultaneously estimated species tree, while also estimating ancestral population sizes and deep coalescence, thereby accommodating ILS (Edwards et al., 2007; Heled & Drummond, 2009). These methods are arguably the most sophisticated in phylogenomics and use a realistic model to explain gene tree incongruence that is not caused by reticulate evolution. However, these methods are only applicable to particular data sets due to the heavy computational load and the need to sample multiple individuals per species. Therefore, other methods have been proposed that rely on first estimating individual gene trees and then either evaluating support and conflict across gene trees with respect to a concatenation-based species tree (Smith et al., 2015), summarizing them within a multi-species coalescent framework (e.g. ASTRAL, Mirarab et al., 2014), inferring (extended) majority-rule consensus trees from them (Salichos & Rokas, 2013), or using Bayesian concordance analysis (Larget et al., 2010) to obtain the most probable species tree topology. A central concept in these methods is that across the phylogeny alternative topologies are evaluated based on the distribution of gene tree relationships, either by estimating quartet support (Sayyari & Mirarab, 2016), internode certainty (Salichos & Rokas, 2013; Kobert et al., 2016) or concordance factors (Larget et al., 2010), comparing the first to the second most prevalent conflicting bipartition or to multiple alternative bipartitions for each node. These methods are likely to become ever more important to disentangle complex evolutionary histories across the ToL as more and more gene trees for additional clades are estimated.

Apart from tackling gene tree incongruence, another notable methodological advance has been the development of more realistic models of sequence evolution, in particular for protein sequences. It has long been acknowledged that poor model fit can lead to inference

artefacts such as long branch attraction (LBA), and when concatenating large numbers of gene alignments, such artefacts are exacerbated. Therefore, it is critical to use adequate substitution models and especially to account for heterogeneity across the alignment caused by different functional and evolutionary constraints across genes, domains and sites. Substitution rate variation among sites is generally accounted for by a discrete gamma model or by partitioning the data by codon position, which is usually extended in supermatrix approaches with partitioning by gene to account for variation between genes. However, these methods fail to adequately capture fine-scale heterogeneity both within and across genes and moreover, for smaller partitions it becomes more challenging to estimate substitution parameters. Therefore, mixture models have been developed to estimate site-specific substitution parameters across a sequence alignment without partitioning the data *a priori*. The CAT (Lartillot & Philippe, 2004) and LG4M/LG4X (Le et al., 2012) models are specifically designed for protein sequences, while the former also gives a good fit to DNA data. Pagel & Meade (2004) described a Bayesian reversible-jump mixture model to detect 'pattern-heterogeneity' in either nucleotide or amino acid sequences. All of these models are able to use the power of the full alignment for estimating substitution parameters by analysing similarly evolving sites regardless of gene or codon position and they have been shown to significantly reduce the occurrence of LBA artefacts (Lartillot & Philippe, 2004; Lartillot et al., 2007).

The other main issue in phylogenomics, studying the mechanisms of molecular evolution, is very broad and includes studies on the evolution of gene function across organisms, gene content and genome size evolution, the process of speciation (including hybridization/introgression) and polyploidization. Originally, the term phylogenomics was introduced to describe a method of functional prediction of genes by evolutionary analysis (Eisen, 1998), but the evolution of gene function remains an important focus of phylogenomics under its current expanded characterization. For example, Lee et al. (2011) used a functional phylogenomic approach to identify gene categories that are associated with the diversification of particular seed plant clades, such as RNA interference genes in monocots. Moore et al. (2017) specifically targeted functionally interesting gene families involved with CAM and C4 photosynthesis and in addition to inferring a species tree, they also

analysed the molecular evolution of these genes in relation to their functions. An example involving legumes is the discovery that non-nodulating legumes still contain genes that are functionally associated with nodulation, providing evidence for multiple losses rather than multiple gains of nodulation (Griesmann et al., 2019).

The process of speciation has also received plenty of attention in phylogenomic studies, for example in relation to lack of resolution caused by rapid successive speciation events and/or hybridization/introgression. For instance, Suh et al. (2015) provided some of the most compelling evidence for ILS by analysing retroposon insertion sites across the radiation of the Neoaves clade of birds. Several studies have used phylogenomic data to study introgression through hybridization, showing that it is common in the evolutionary history of several taxonomic groups (e.g. Marcet-Houben & Gabaldón, 2015; Li et al., 2016; Meier et al., 2017), including many examples in angiosperms, as long suspected (e.g. Eaton & Ree, 2013; Escudero et al., 2014; Folk et al., 2017; Morales-Briones et al., 2018). Rapid radiations, such as those of cichlid fishes (Brawand et al., 2014), Galapagos finches (Lamichhaney et al., 2015) and wild tomatoes (Pease et al., 2016), have been subjected to various types of phylogenomic analysis, highlighting several interesting aspects of how speciation proceeds during radiations, including the sorting of ancestral variation, *de novo* gene origination and accelerated coding sequence evolution.

Polyploidization (or WGD) is a recurrent feature of genome evolution and ancient WGD events are thought to have been important in the evolution of several major groups, including yeasts (Wolfe & Shields, 1997; Kellis et al., 2004) tetrapods (Dehal & Boore, 1005), teleost fishes (Glasauer & Neuhauss, 2014), frogs (Session et al., 2016), and above all across plants, e.g. seed plants and angiosperms (Jiao et al., 2011), monocots (Jiao et al., 2014), Pentapetalae (Jiao et al., 2012), Caryophyllales (Yang et al, 2018), Malphigiales (Cai et al., 2019), Brassicaceae (Huang et al., 2015), Asteraceae (Barker et al., 2016; Huang et al., 2016), Malvaceae (Conover et al., 2019) and Leguminosae (Cannon et al., 2015; Chapter II of this thesis). Phylogenomics offers unprecedented opportunities to study (ancient) polyploidy and has been revealing just how prevalent and widespread ancient WGDs were throughout the evolutionary history of angiosperms in particular (e.g. Yang et al., 2018; Cai et al., 2019). One of the most interesting findings is that ancient WGDs appear to be significantly

concentrated around the Cretaceous-Paleogene (K-Pg) boundary (Fawcett et al., 2009; Vanneste et al., 2014; Lohaus & Van de Peer, 2016), as extensively discussed in Chapter II. Ancient polyploidy also occurred during the early evolution of the legume family, and Cannon et al. (2015) and Stai et al. (2019) inferred as many as five independent non-nested polyploidy events occurred, rather than a single shared WGD, rendering most but not all extant legumes paleopolyploid. In Chapter II, these findings of independent WGDs and their phylogenetic placements are evaluated using gene family trees that include five of the six major lineages of legumes and the hypothesis that the origin of the legumes and legume WGD events are linked to the K-Pg boundary is tested.

Several methods to discover and infer the phylogenetic placement of WGDs have been proposed. When high-quality genome assemblies are available, ancient WGDs can be detected from colinearity of duplicated syntenic blocks, which generally provides strong evidence for WGDs. However, the drawback is that high-quality assemblies are still only available for a relatively small number of plant species. For example in legumes, while “draft genomes” have recently been published for *Cercis* (Cercidoideae), *Chamaecrista* and *Mimosa* (Caesalpinioideae) (Griesmann et al., 2019), high-quality assemblies are only available for crop legumes and their close relatives within the 50Kb-inversion clade of Papilionoideae. Indeed, a WGD shared among Papilionoids was first discovered using the synteny approach to compare the genomes of *Glycine*, *Medicago* and *Lotus* (Mudge et al., 2005; Cannon et al., 2006). Another commonly used approach is to estimate the synonymous distance (K_s) between all paralog pairs in a genome, and plot the resulting distribution of these distances. Since an ancient WGD will have led to simultaneous divergence of multiple paralog pairs, these events are visible as peaks against the background of small-scale duplication events. However, the *Ks* plot method has serious shortcomings (Tiley et al., 2018), meaning that it cannot reliably detect WGDs when these are very old due to saturation of synonymous substitutions and it also fails when ≤10% of paralog copies are retained. Most of the evidence for multiple independent WGDs in legumes stems from K_s plot analyses, so additional evidence for the phylogenetic placement of these is necessary. Phylogenomic methods to study ancient WGDs (i.e. those that take phylogenetic information across gene families into account) have also been proposed with many recent developments to address

this question. These methods rely on the presence of multiple WGD-derived paralog copies across gene family trees, which are reconciled with the species tree. These gene tree reconciliation approaches can be relatively simple, counting numbers of descendent duplicates for each node as in the program Phyparts (Smith et al., 2015) or complex, involving models of gene duplication and loss (often augmented by modeling ILS, LGT and/or introgression, e.g. using the programs Notung (Stolzer et al., 2012), GRAMPA (Gregg et al., 2017) and WHALE (Zwanepoel & Van De Peer, 2019)). A probabilistic approach that uses gene count data from gene families and a species tree topology has also been proposed (Rabier et al., 2013), providing a powerful tool for hypothesis testing, although it does not take gene tree topologies into account. In fact, all gene tree reconciliation methods have limitations, meaning that combining multiple different methods, dependent on the particularity of WGD in the evolutionary history of the clade under study, will be necessary in most cases to obtain sufficient evidence for the occurrence and phylogenetic placement of WGDs.

Apart from the two main issues, other related fields in evolutionary biology are also starting to be transformed by the use of phylogenomic data. For example, in macro-evolutionary studies that use time-calibrated phylogenies to infer timings of diversification and biogeography, genome-scale data can potentially infer more accurate divergence time estimates since branch lengths are more accurately estimated. However, larger data sets may merely increase precision, while the idiosyncrasies of the molecular clock and possible inadequacies of calibration priors are not overcome by using more data. As a consequence, phylogenomic dating has not ended controversies on timings of origin of major taxonomic groups. For instance, time-calibration analysis based on 1,478 protein-coding genes to infer the timing of insect evolution has led to age estimates that are mostly congruent with the appearance of insect lineages in the fossil record (Misof et al., 2014), but this has been questioned in particular for the use of overly informative priors and the placement of fossils (Tong et al., 2015). Also in vertebrates, controversy remains over the age estimates of major groups, in particular in relation to the K-Pg boundary, for example in mammals and birds, as discussed further in Chapter II. A recent study thoroughly analysed 7,189 nuclear genes to estimate a timetree for the jawed vertebrates (Irisarri et al., 2017), but ignored the effect that the marginal priors can have on divergence time estimation (Brown & Smith, 2017), while also

only selectively citing other studies that found similar ages, thereby disregarding controversy surrounding age estimates in vertebrates. The age of the angiosperms is perhaps the most controversial molecular dating problem (see Magallon et al., 2015; Beaulieu et al., 2015; Herendeen et al., 2017). A recent study that used 80 chloroplast genes finds a Triassic origin of the angiosperms (Li et al., 2019), as have several previous studies based on much less data (see Magallon et al., 2015). However, the method used (penalized likelihood using bootstrap replicate trees), undoubtedly chosen by Li et al. because of the large tree size, is now considered inferior to more recently developed Bayesian approaches to divergence time estimation. Moreover, the chloroplast genome has been questioned as a reliable source of data for divergence time estimation due to more pronounced substitution rate heterogeneity relative to nuclear genes (Christin et al., 2014). Therefore, the “Jurassic gap” that Li et al. (2019) described, indicating the absence of angiosperm fossils in the Jurassic, may not be real, but could rather be caused by the use of unsuitable data and methods, as a Triassic age of angiosperms is not supported by paleobotanical evidence (Coiro et al., 2019).

The methods to infer timetrees using genome-scale data are mostly identical to analyses that use only one or a few gene sequences, except that topological uncertainty is usually not taken into account due to the heavy computational cost involved, and analyses are instead performed on a fixed topology, justified by the fact that species trees inferred from genome-scale data are more robust. Even then, the volumes of data can cause high computational loads. Therefore, gene-jackknifing methods have been used (Irisarri et al., 2017) and a method to select relatively informative and clock-like genes to reduce data set size has been proposed (Smith et al., 2018).

The Leguminosae

"The family Leguminosae, though it ranks behind the Orchidaceae and the Compositae in total number of genera and species, offers by all other criteria the most spectacular example of evolutionary and ecological radiation of any angiosperm family. ... In their diversity of life forms, life histories, habits, geographical distributions, and relationships with pollinators, seed

dispersers, herbivores, and other animal associates, the legumes are unmatched by any other family." ***- McKey, 1994.***

This quote captures very well the main reasons why legumes are an exceptionally interesting group of plants and an excellent study group for evolutionary research. The large morphological and ecological diversity of legumes means that the family offers opportunities to answer specific questions on a great variety of aspects of flowering plant evolution. Moreover, because legumes are omnipresent in most vegetation types across the globe, together with the large number of species, the study of legume evolution has the potential to answer big and general questions about the evolution of plant diversity.

The family Leguminosae was first established by Adanson in 1763 in his *Familles des Plantes*. It was subdivided by De Candolle (1825) into four groups, which was elaborated upon by Bentham (1865), who recognised three major groups that have since been recognized variously at the family or subfamily rank until recently, when LPWG (2017) proposed a new phylogenetic subfamily classification involving six subfamilies. 770 legume genera are currently recognised (ranging from 1 to c. 503 per subfamily), in total harbouring c. 19,581 described species (ranging from 1 to c. 14,000 per subfamily). The steady accumulation of species, with publication of twenty to thirty species new to science each year, suggests that the total number of extant legume species, including those that remain to be discovered and described, surpasses the 20,000 species mark, just like Orchidaceae and Compositae (or Asteraceae), the only two plant families that exceed the number of legume species.

Legumes are either trees, shrubs, lianas, suffrutices, herbs or vines, covering nearly the full spectrum of angiosperm habits. They typically have alternate compound leaves, although unifoliolate leaves can also be found in particular in Cercidoideae and Papilionoideae, with opposite or whorled leaves also found occasionally. Stipules are usually present, sometimes foliaceous. Inflorescences are highly diverse, although often racemes, but paniculate, globose and spicate inflorescences are common as well. Floral diversity in the family is truly exceptional (Fig. 1). While the majority of species in the largest subfamily is characterized by pea-shaped (or papilionate) flowers, in all subfamilies (except the monotypic

Figure 1. Flower diversity in the Leguminosae. From left to right, top row: *Chadsia longidentata*, *Vachellia karroo*, *Bauhinia madagascariensis* and *Ornithopus perpusillus*, second row: *Bussea sakalava*, *Campsiandra comosa*, *Anthyllis tetraphylla* and *Senna leandrii*, third row: *Baudouinia* sp., *Paramacrolobium coeruleum*, *Calliandra fuscipila* and *Lupinus cosentinii*, bottom row: *Delonix pumila*, *Pisum sativum*, *Medicago marina* and *Entada chrysostachys.* All photos by Erik Koenen.

Duparquetioideae) floral morpholgy is far from homogeneous. Flowers can be radially or bilaterally symmetrical or rarely asymmetrical and the number of sepals, petals and stamens is variable and these can be free or fused, equal or unequal. While usually bisexual, unisexual flowers can also be found and stamens can be modified into staminodes in sometimes sterile flowers. Finally, the defining feature of the family, the fruit (often referred to as the ‘legume’ or ‘pod’) also shows remarkable morphological diversity, although it always consists of a single superior carpel with a unilocular ovary and has parietal placentation in two alternating rows.

Building on this basic ground plan, legume fruits have evolved various modes of dehiscence or are often indehiscent, and can then sometimes be a lomentum, samara or drupe. Fruits may contain a single to many seeds, which can be coloured and with or without a pleurogram, fleshy aril, sarcotesta and/or wings. (Information from Lewis et al. (2005) and LPWG (2017)).

The large morphological diversity of legumes probably reflects the adaptive radiation of the family across nearly all vegetation types across nearly the whole planet, Antarctica being the only terrestrial region where legumes are absent. Legumes are not well-represented in tropical montane forests, and are not found in mangroves, but they are diverse and dominant in many lowland tropical forests, grasslands and semi-arid regions as well as in temperate, alpine and arctic grasslands and tundra and Mediterranean-type climatic regions. They are often woody in the tropics, usually herbaceous in the temperate zone. All legumes are further characterized by high nitrogen content in the leaves, and most members of Caesalpinioideae and Papilionoideae form a symbiotic relationship with nitrogen-fixing ‘rhizobia’ bacteria in root nodules (McKey et al., 1994; Sprent & Platzmann, 2001). Given the importance of nodulation both in an ecological and agricultural context, this trait has been intensively studied. Long thought to have evolved independently in several lineages of legumes from a shared cryptic precursor trait, nodulation has recently been shown to be instead lost massively and in parallel (Griesmann et al., 2019) not only in legumes but across the nitrogen-fixing clade of angiosperms as a whole (van Velzen et al., 2019).

Legumes most likely first evolved in the late Cretaceous or earliest Paleocene, since no uncontroversial legume fossils pre-date the K-Pg boundary. However, this has not been convincingly tested using molecular divergence time estimation, and the hypothesis that the crown node of the legumes is associated with the K-Pg boundary is tested in Chapter II. The biogeographical origin of the legumes also remains unclear, and is challenging to reconstruct for such a large and widely distributed family. Schrire et al. (2005) performed the most comprehensive ancestral area reconstruction and suggested that legumes originated around the Tethys Seaway in seasonally dry tropical vegetation. However, considering that the oldest fossil evidence of legumes stems from Patagonia (Brea et al., 2008), Schrire’s hypothesis seems unlikely unless these fossils represent stem-relatives of the family. Moreover, when studying the geographic distributions of successive sister-lineages to the core clades of each

subfamily, it becomes clear that these generally occupy disparate regions on the planet. For instance, the genera *Umtiza* (1 sp., Southern Africa), *Gleditsia* (c. 13 spp., East Asia and the Americas), *Gymnocladus* (6 spp., East Asia and North America), *Tetrapterocarpon* (2 spp., Madagascar), *Arcoa* (1 sp., Caribbean), *Acrocarpus* (1 sp., India and South East Asia) and *Ceratonia* (2 spp., North East Africa and the Mediterranean), form two successive species-poor sister clades to the core Caesalpinioideae clade that contains the remaining c. 4,374 species in that subfamily. Similar patterns are apparent in Papilionoideae, Dialioideae, Detarioideae and Cercidoideae. These depauperate lineages with deep-branching stem origins may constitute relicts from more diverse and widespread initial radiations of these subfamilies (i.e. “dying embers” sensu Spriggs et al., 2015), or they may have originated elsewhere, with dispersal to their current ranges theoretically taking place any time along their long stem lineages. Either way, the biogeographic origins of these subfamilies will be obscured in any phylogeny inferred solely from extant taxa, as will that of the family, for which fossil evidence provides a South American origin as the best guess. In Paleocene fossil sites in Colombia (Wing et al., 2009; Herrera et al., in press) and Eocene deposits in North America, Europe, Asia and Africa (Herendeen & Dilcher, 1992), legumes are common and highly diverse, and remain diverse in later fossil sites across several continents throughout the Oligocene and Neogene. This suggests that legumes are not only dominant and diverse across the planet at present, but that they have been so throughout most of the Cenozoic, in both the warm Paleogene and the cooler Neogene, further highlighting the adaptive potential of the family.

The drivers of legume diversification remain poorly understood. Koenen & de Vos et al. (2013) estimated diversification rates across the family and in greater detail for several well-sampled legume clades, and showed that diversification rates vary widely across the family. However, as to the causes of higher diversification in some clades relative to others, no general pattern was observed. Some of the variation in observed diversification rates across the family is suggested to be related to distinct ecologies, with distinct trajectories of diversification in different biomes. However, for several fast evolving legume clades, no clear feature that can explain the higher rates is apparent, in line with the idea that the triggering of rapid radiations is highly complex (Bouchenak-Khelladi et al., 2015). On the other hand, more

detailed studies on clades with well-sampled phylogenies have led to suggestions that particular trait and/or ecological shifts can explain higher diversification rates, such as a shift to high elevation vegetation coupled with the evolution of perennial habit in *Lupinus* (Hughes & Eastwood, 2006; Drummond et al., 2012; Hughes & Atchison, 2015), the evolution of extra-floral nectaries in *Senna* (Marazzi & Sanderson, 2010) and interactions with herbivores in *Inga* (Richardson et al., 2001; Kursar et al., 2009). Similar ecological shifts in *Calliandra*, *Mimosa* and *Indigofera,* respectively to Brazilian campos rupestres, the Cerrado and South-African fynbos, are also found to be associated with higher diversification rates (Koenen & de Vos et al., 2013). However, Koenen & de Vos et al. (2013) noted that their analyses suffered from several methodological limitations, of which the unreliable estimation of extinction rates from phylogenies is perhaps the most problematic. Indeed, their study did not shed much light on the origin of high Eocene diversity of legumes apparent in the fossil record. Since most of the clades that show increased diversification appear to post-date the Eocene, this suggests that much of the Paleogene diversity has gone extinct, leading to evolutionary turnover of lineages (Koenen et al., 2015), likely caused by the cooling trend that followed the Eocene climatic optimum and continued until recently (Zachos et al., 2001).

The mimosoid clade

In Chapter III, I focus on the mimosoid clade (formerly Mimosoideae) a subclade of subfamily Caesalpinioideae. This clade constitutes a highly characteristic group of legumes, comprising c. 3,300 species of trees, shrubs, suffrutices, lianas and a few herbaceous aquatics (*Neptunia*), in c. 87 genera. Some of the diversity of inflorescences, fruits and leaves in mimosoids is depicted in Figures 2-4. The clade is most diverse in the tropics and subtropics, with few species extending to temperate regions in the Northern Hemisphere, extratropical South America and Southern Africa, and two moderately diverse assemblages of *Acacia* s.s. in the temperate SW and SE corners of Australia. Across the lowland tropics, the clade is found in nearly all biomes and vegetation types (Fig. 5), except mangroves. Mimosoids are highly abundant in rainforests in the Americas and Africa, form prominent or even dominant elements of the woody floras of tropical grasslands (savannas) in Brazil,

Figure 2. Inflorescence diversity of mimosoids. From lef to right, top row: *Macrosamanea amplissima*, *Inga subnuda* and *Mimosa volubilis* directly beneath, heteromorphic inflorescence of *Dichrostachys cinerea, Senegalia ataxacantha*, bottom row: dimorphic inflorescences of *Hydrochorea corymbosa* and *Albizia grandibracteata*, heteromorphic inflorescence of *Parkia bahiae*, *Acacia* sp. All photos by Erik Koenen.

across Africa and in Australia, and dominate seasonally dry tropical forests (SDTFs sensu Pennington; the succulent biome sensu Schrire et al. 2005 & Ringelberg et al., submitted) in Central America, the Caribbean, North-East Brazil, the Horn of Africa and Madagascar. The South-East Asian tropics are less rich in mimosoids, as is true for legumes more generally, but some groups of mimosoids, such as the genera *Adenanthera*, *Albizia* and *Archidendron* and allies, did diversify in this region as well, , the latter also extending into the Pacific region.

Figure 3. Fruit diversity of mimosoids. From left to right, top row: *Hydrochorea emarginata, Enterolobium contortisiliquum* and below it *Inga subnuda, Newtonia hildebrandtii, Calliandra viscidula* and below it *Pentaclethra macroloba* and *Chloroleucon foliolosum*, bottom row: *Plathymenia reticulata*, *Abarema cochliacarpos, Macrosamanea amplissima*. All photos by Erik Koenen.

Given their ubiquity across the tropics, the mimosoids offer opportunities to study macro-evolutionary diversification on a truly pantropical scale.

The mimosoids are a morphologically distinctive group of legumes, with radially symmetrical flowers with valvate petal and sepal aestivation, and in most species with prominent, often coloured stamens for pollinator attraction, and a majority of species is characterized by bipinnate leaves (the large genera *Inga* and *Acacia* s.s. with once-pinnate leaves or leaves modified into phyllodes, respectively, being the main exceptions; Fig. 4).

Figure 4. Leaf diversity of mimosoids. From left to right, top row: *Zygia cataractae*, *Inga pilosula*, *Acacia longifolia*, bottom row: *Macrosamanea amplissima, Albizia tanganyicensis* and below it *Abarema barbouriana* var. *arenaria*, *Vachellia sieberiana.* All photos by Erik Koenen.

Given this distinctiveness, the clade has long been considered one of the three main groups of legumes, previously classified as a subfamily or even a separate family. However, radially symmetrical flowers are also present in several other lineages of Caesalpinioideae (sensu LPWG, 2017) as well as in other subfamilies. These non-mimosoid Caesalpinioids with radially symmetrical flowers have imbricate petal aestivation, but so does the mimosoid genus *Parkia*, which is deeply nested within the mimosoid clade. Bipinnate leaves are also rather common across the other non-mimosoid lineages of Caesalpinioideae. Taken together, this suggests that mimosoids are not as distinctive as previously thought and their inclusion in

Figure 5. Mimosoid habitats. From left to right, top row: Semi-arid desert, seasonally dry tropical scrub, Cerrado fire-prone grassland, “Igapo” flooded Amazonian forest, bottom row: seasonally dry tropical scrub, African savanna, submontane rainforest and lowland terra firme rainforest. Top row first two photos from the left and photo in the bottom left corner by Colin Hughes, other photos by Erik Koenen.

Caesalpinioideae is therefore well justified (LPWG, 2017). A new tribal and/or clade-based classification is needed for the recircumscribed Caesalpinioideae including the mimosoid clade, as discussed in Chapter III. Species diversity is rather unequally distributed across the mimosoid clade. More than half of all mimosoid species are included in just three genera: *Acacia* s.s. (c. 1000 spp., ref), *Mimosa* (c. 550 spp., Barneby; Simon) and *Inga* (c. 300 spp.). Furthermore, the mimosoid generic backbone phylogeny is strongly imbalanced, and species diversity is mainly concentrated in the Ingioid clade (> 2000 spp.) and in the genus *Mimosa*.

The Ingioid clade has been notoriously problematic in terms of generic delimitation (Brown, 2008) and has remained largely unresolved in current phylogenies. In Chapter III, the problems surrounding Ingioid classification are discussed and results from phylogenetic analyses of hybrid capture data are presented and discussed in relation to (lack of) phylogenetic resolution in the Ingioid clade.

Several mimosoid traits may have facilitated their pantropical adaptive radiation, such as spinescence (Fig. 6), extra-floral nectaries (EFNs; Fig. 6), pollen in polyads and nodulation. Spines and prickles can have several functions, including anti-herbivory defense, avoidance of excessive water loss and to mount surrounding vegetation in species with a climbing or scrambling habit. EFNs are involved in attracting ants, which will defend their host plant against insect herbivores. Apart from EFNs, some mimosoids in the genus *Vachellia* have evolved other traits, such as domatia in enlarged and swollen stipular spines and the so-called Beltian bodies, modified leaflet tips that are consumed by the ants and are rich in lipids, sugars and proteins (Fig. 6) for the recruitment of ants in anti-herbivory defense. Many mimosoids have their pollen aggregated into polyads. These polyads are similar to pollinia in orchids and milkweeds and other forms of pollen aggregation in various lineages, the evolution of which appears complex but probably related to more efficient siring of all ovules in a fruit (Harder & Johnson, 2008). This is likely to increase seed set following successful cross-pollination (Seijo & Solis Neffa, 2004). Nitrogen-fixation by rhizobia in root nodules is commonly found across most species of mimosoids, which may help them to maintain the nitrogen-demanding lifestyle typical of all legumes, leading to the benefit of highly productive nitrogen-rich leaves when photosynthesis is not limited in any other way (McKey, 1994). Therefore, this trait may be important for achieving dominance, as well as for colonizer species which are commonly found among mimosoids, as it enables rapid vigorous growth. Moreover, highly productive nitrogen-rich leaves may be particular advantageous in seasonally dry regions, where the ability to quickly photosynthesise during favourable periods and shed the leaves at the start of a period of drought (McKey, 1994), may further explain the dominance of mimosoids in the savanna and succulent biomes.

Figure 6. Mimosoid functional traits. From left to right, top row: extra-floral nectaries on *Inga edulis, Stryphnodendron rotundifolium* and *Macrosamanea amplissima*, bottom row: Stipular spines in *Prosopis ferox,* modified stipular spines with ant domatia and beltian bodies at the leaflet tips of *Vachellia cornigera*. All photos by Erik Koenen, except in the bottom left corner, photo by Colin Hughes.

Important as these traits may be in providing the background for and/or triggering the mimosoid radiation (Bouchenak-Khelladi et al., 2015), these traits are rather uniform across the taxa that possess them. Traits that are highly and flexibly evolvable may be more important in generating large numbers of species as they are likely the traits involved with niche-partitioning across the available ecological space. Such traits are referred to as ‘modulators’ by Bouchenak-Khelladi et al. (2015) and they constitute traits that are variously modified across a species radiation. Key candidate modulator traits in mimosoids are the

compound inflorescence, seed dispersal syndrome and bipinnate leaf. These three traits are highly variable across the clade and therefore likely to be involved with adaptation to different environments. The various ways in which the mostly globose or spicate inflorescences of mimosoids are aggregated into so-called compound inflorescences, as well as various modifications of different flowers within the same inflorescence unit (leading to dimorphic or heteromorphic inflorescences, Fig. 2), likely facilitated the recruitment of local pollinators when adapting to new environments. Similarly, seed dispersal that can occur by explosive dehiscence (ballistic dispersal or ballochory), zoochory, hydrochory or anemochory, with considerable variation within these categories, appear linked to adaptation to the local settings of e.g. the rainforests of the Amazon and Congo basins (where zoochory and hydrochory are common) or more open and seasonally dry habitats (zoochory, ballochory and anemochory are common in deciduous forests and grasslands across tropical America and Africa). However, while inflorescence and fruit morphology are often rather uniform within genera, leaf traits appear to be the most variable trait across the mimosoid clade (Fig. 4). Similar variation in the size and shape of leaflets as well as the number of leaflets and pinnae can be observed among and within both more and less closely related genera, suggesting that these wide ranges of leaf dimensions have evolved numerous times in numerous different lineages of mimosoids. The number of leaflets per leaf, for example, ranges from several thousand individual leaflets on several tens of pinnae pairs in finely divided leaves (e.g. in *Parkia* or *Vachellia*) to the extreme case in which a bifoliolate, yet still bipinnate leaf is formed by two pinnae with one leaf each (in *Zygia unifoliolata*). In others, leaflets are absent and the petiole and rachis are modified into phyllodes, in particular in the genus *Acacia* in Australia, while some species in semi-arid regions are completely aphyllous and photosynthesis occurs in green stems (e.g. in some *Prosopis* and *Acacia* s.s.). Generally, in tropical humid forests, leaves and/or leaflets are large and evergreen, while in seasonally dry regions they are nearly always deciduous and often finely divided, and relatively small to minute. This tremendous disparification of leaf morphology likely underlies high functional diversity in relation to adaptation along the tropical wet-dry and seasonality climatic gradients. The idea that this trait in particular is key to the evolutionary success of the mimosoids is a critical hypothesis to test in future research when a well-resolved and well-sampled phylogeny

of the clade is available. Chapter III presents a new protocol for genome-scale data generation and phylogeny inference in mimosoids, and paves the way for future reclassification as well as macroevolutionary studies of the clade and adaptation to the different climatic regions in the global tropics.

Main objectives and description of the chapters

In this thesis, phylogenomic methods are applied to study the early evolution of the legume family and the prominent mimosoid clade of subfamily Caesalpinioideae. The main objectives are:

- To resolve the earliest dichotomies within the legume phylogeny and the relationships among subfamilies (Chapter I)
- To infer an enhanced phylogeny for the mimosoid clade, and in particular to resolve the Ingioid clade (Chapter III)
- To understand which evolutionary processes and/or methodological limitations underlie the lack of resolution observed in parts of the legume phylogeny (all chapters)
- To disentangle the complex history of paleopolyploidy during the early evolution of the legumes (Chapter II)
- To infer the timings of the origin of the legumes and ancient WGD events (Chapter II)

In Chapter I, the assembly and analysis of phylogenomic data sets derived from plastome sequences, completely sequenced genomes and transcriptomes are described. The main aim of the study is to infer a new phylogenetic framework for legume evolution using genome-scale data, in particular to improve phylogenetic resolution across the first divergences in the family, which have proven difficult to resolve. In order to assess the robustness of the inferred species tree and shed light on why these relationships have been difficult to resolve, the strength of phylogenetic signal across gene trees and support for alternative relationships are evaluated.

In Chapter II, the nuclear genomic data set generated for Chapter I is used to study the history of polyploidy during the early evolution of the legumes and to estimate the timing of WGDs in relation to the origin and early diversification of the family and the K-Pg boundary. The hypotheses that are tested in this chapter are: (1) the opposing hypotheses of shared paleopolyploidy among several or all subfamilies with potentially multiple nested WGDs, versus independent and non-nested WGD events in each subfamily; and (2) that ancient polyploidy and the initial diversification of the legume family are associated to the K-Pg boundary.

Chapter III focuses on the mimosoid legumes, a former subfamily and prominent clade within Caesalpinioideae, and attempts to answer the following questions: (1) can targeted sequencing through hybrid capture of low-copy nuclear genes provide useful data to infer an enhanced phylogeny for the mimosoid clade?; (2) have previous classification schemes for the Ingioid clade and generic delimitation within these led to the recognition of natural, monophyletic groups, and if not, can we infer robustly supported subclades within the Ingioids?; and (3) what has caused the difficulties in obtaining resolution within the Ingioid clade?

References

Adanson, M., 1763. Familles des plantes, vol. 2. Paris: chez Vincent.

Angiosperm Phylogeny Group, 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden*, 85(4):531-553.

Angiosperm Phylogeny Group, 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical journal of the Linnean Society*, 141(4):399-436.

Angiosperm Phylogeny Group, 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, 161(2):105-121.

Angiosperm Phylogeny Group, Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W., Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S. and Stevens, P.F.,

2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1-20.

Atchison, G.W., Nevado, B., Eastwood, R.J., Contreras-Ortiz, N., Reynel, C., Madriñán, S., Filatov, D.A. and Hughes, C.E., 2016. Lost crops of the Incas: Origins of domestication of the Andean pulse crop tarwi, *Lupinus mutabilis*. *American Journal of Botany*, 103(9):1592-1606.

Bapteste, E., O'Malley, M.A., Beiko, R.G., Ereshefsky, M., Gogarten, J.P., Franklin-Hall, L., Lapointe, F.J., Dupré, J., Dagan, T., Boucher, Y. and Martin, W., 2009. Prokaryotic evolution and the tree of life are two different things. *Biology Direct*, 4(1):34.

Barker, M.S., Li, Z., Kidder, T.I., Reardon, C.R., Lai, Z., Oliveira, L.O., Scascitelli, M., Rieseberg, L.H., 2016. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *American Journal of Botany*, 103:1203–1211.

Beaulieu, J.M., O'Meara, B.C., Crane, P., and Donoghue, M.J., 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Systematic Biology*, 64(5):869-878.

Bentham, G., 1865. Leguminosae. In: Bentham, G. & Hooker, J.D. (eds.), Genera plantarum, vol. 1(2). Londini [London]: venit apud Lovell Reeve. Pp. 434–600.

Bouchenak-Khelladi, Y., Onstein, R.E., Xing, Y., Schwery, O. and Linder, H.P., 2015. On the complexity of triggering evolutionary radiations. *New Phytologist*, 207(2):313-326.

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E. and Turner-Maier, J., 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):375.

Brea, M., Zamuner, A.B., Matheos, S.D., Iglesias, A., Zucol, A.F., 2008. Fossil wood of the Mimosoideae from the early Paleocene of Patagonia, Argentina. *Alcheringa*, 32:427–441.

Brown, G.K., 2008. Systematics of the tribe Ingeae (Leguminosae-Mimosoideae) over the past 25 years. *Muelleria*, 26(1):27-42.

Brown, J.W. and Smith, S.A., 2017. The past sure is tense: on interpreting phylogenetic divergence time estimates. *Systematic Biology*, 67(2):340-353.

Cai, L., Xi, Z., Amorim, A.M., Sugumaran, M., Rest, J.S., Liu, L. and Davis, C.C., 2019. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytologist*, 221: 565-576.

Candolle, A. de, 1825. Prodromus systematis naturalis regni vegetabilis, vol. 2. Parisiis [Paris]: sumptibus sociorum Treuttel et Würtz.

Cannon, S.B., Sterc, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., Schiex, T., Spannagl, M., 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proceedings of the National Academy of Sciences, USA*, 103:14959–14964.

Cannon, S.B., McKain, M.R., Harkess, A., Nelson, M.N., Dash, S., Deyholos, M.K., Peng, Y., Joyce, B., Stewart Jr, C.N., Rolf, M., Kutchan, T., 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution*, 32(1):193–210.

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A.M., Bittrich, V., Celis, M., Daly, D.C., Fiaschi, P., Funk, V.A., Giacomin, L.L. and Goldenberg, R., 2017. Amazon plant diversity revealed by a taxonomically verified species list. *Proceedings of the National Academy of Sciences, USA*, 114(40):10695-10700.

Cavender-Bares, J., González-Rodríguez, A., Eaton, D.A., Hipp, A.A., Beulke, A. and Manos, P.S., 2015. Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): a genomic and population genetics approach. *Molecular Ecology*, 24(14):3668-3687.

Christin, P.A., Edwards, E.J., Besnard, G., Boxall, S.F., Gregory, R., Kellogg, E.A., Hartwell, J. and Osborne, C.P., 2012. Adaptive evolution of C4 photosynthesis through recurrent lateral gene transfer. *Current Biology*, 22(5):445-449.

Christin, P.A., Spriggs, E., Osborne, C.P., Strömberg, C.A.E., Salamin, N., Edwards, E.J., 2014. Molecular dating, evolutionary rates, and the age of the grasses. S*ystematic Biology*, 63:153–165.

Coiro, M., Doyle, J.A. and Hilton, J., 2019. How deep is the conflict between molecular and fossil evidence on the age of angiosperms?. *New Phytologist*, 223:83–99.

Conover, J.L., Karimi, N., Stenz, N., Ané, C., Grover, C.E. Skema, C., Tate, J.A., Wolff, K., Logan, S.A., Wendel, J.F., Baum, D.A., 2019. A Malvaceae mystery: A mallow maelstrom of genome multiplications and maybe misleading methods? *Journal of Integrative Plant Biology*, 61: 12-31.

Couvreur, T.L., Helmstetter, A.J., Koenen, E.J., Bethune, K., Brandão, R.D., Little, S.A., Sauquet, H. and Erkens, R.H., 2018. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. *Frontiers in Plant Science*, 9:1941.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V. and Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, 99(2):291-311.

Darwin, C., 1859. On the Origin of Species. London: John Murray

de Sousa, F., Foster, P.G., Donoghue:C., Schneider, H. and Cox, C.J., 2019. Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist*, 222(1):565-575.

Dehal, P. and Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10):e314.

Delsuc, F., Brinkmann, H. and Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361.

Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124-2128.

Doolittle, W.F. and Brunet, T.D., 2016. What is the tree of life?. *PLoS Genetics*, 12(4):e1005912.

Drummond, C.S., Eastwood, R.J., Miotto, S.T. and Hughes, C.E., 2012. Multiple continental radiations and correlates of diversification in *Lupinus* (Leguminosae): testing for key innovation with incomplete taxon sampling. *Systematic Biology*, 61(3):443-460.

Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J., Yang, Y., Dionora, J., Quick, W.P., Park, M., Bennetzen, J.L. and Besnard, G., 2019. Lateral

transfers of large DNA fragments spread functional genes among grasses. *Proceedings of the National Academy of Sciences, USA*, 116(10):4416-4425.

Eaton, D.A. and Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*, 62(5):689-706.

Edwards, S.V., Liu, L. and Pearl, D.K., 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences, USA*, 104(14):5936-5941.

Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8(3):163-167.

Escudero, M., Eaton, D.A., Hahn, M. and Hipp, A.L., 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in Carex (Cyperaceae). *Molecular Phylogenetics and Evolution*, 79:359-367.

Fawcett, J.A., Maere, S., Van de Peer, Y., 2009. Plants with double genomes might have had a better chance to survive the Cretaceous – Tertiary extinction event. *Proceedings of the National Academy of Sciences, USA*, 106:5737–5742.

Folk, R.A., Mandel, J.R. and Freudenstein, J.V., 2017. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Systematic Biology*, 66(3):320-337.

Glasauer, S.M. and Neuhauss, S.C., 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics*, 289(6):1045-1060.

Gregg, W.T., Ather, S.H. and Hahn, M.W., 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Systematic Biology*, 66(6):1007-1018.

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B., Billault-Penneteau, B., Lauressergues, D., Keller, J., Imanishi, L. and Roswanjaya, Y.P., 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*, 361(6398):eaat1743.

Grover, C.E., Salmon, A. and Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, 99(2):312-319.

Harder, L.D. and Johnson, S.D., 2008. Function and evolution of aggregated pollen in angiosperms. *International Journal of Plant Sciences*, 169(1):59-78.

Heled, J. and Drummond, A.J., 2009. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570-580.

Hennig, W., 1966. Phylogenetic Systematics. Urbana: University of Illinois Press

Herendeen, P.S., Dilcher, D.L., 1992. Advances in legume systematics part 4. The fossil record. Richmond, UK: Royal Botanic Gardens, Kew.

Herendeen, P.S., Friis, E.M., Pedersen, K.R. and Crane, P.R., 2017. Palaeobotanical redux: revisiting the age of the angiosperms. *Nature Plants*, 3(3):17015.

Herrera, F., Carvalho, M.R., Wing, S.L., Jaramillo, C., Herendeen, P.S., 2019. Middle to Late Paleocene Leguminosae fruits and leaves from Colombia. *Australian Systematic Botany*. In press.

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch ,M.A., Al-Shehbaz, I., Edger, P.P., Pires, J.C., Tan, D.-Y., Zhong, Y., Ma, H., 2015. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, 33:394–412.

Huang, C.-H., Zang, C., Liu, M., Hu, Y., Gao, T., Qi, J., Ma, H., 2016. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Molecular Biology and Evolution*, 33:2820–2835.

Hughes, C. and Eastwood, R., 2006. Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proceedings of the National Academy of Sciences, USA*, 103(27):10334-10339.

Hughes, C.E. and Atchison, G.W., 2015. The ubiquity of alpine plant radiations: from the Andes to the Hengduan Mountains. *New Phytologist*, 207(2):275-282.

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M. and Philippe, H., 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology & Evolution*, 1(9):1370.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S. and Soltis, D.E., 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J. and Wu, X., 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology*, 13(1):R3.

Jiao, Y., Li, J., Tang, H. and Paterson, A.H., 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell*, 26(7):2792-2802.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T. and Leebens-Mack, J.H., 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*, 68(4):594-606.

Kellis, M., Birren, B.W. and Lander, E.S., 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617.

Kew. 2016. State of the world's plants. Available from https://stateoftheworldsplants.com/2016/.

Kobert, K., Salichos, L., Rokas, A. and Stamatakis, A., 2016. Computing the internode certainty and related measures from partial gene trees. *Molecular Biology and Evolution*, 33(6):1606-1617.

Koenen, E.J.M., De Vos, J.M., Atchison, G.W., Simon, M.F., Schrire, B.D., De Souza, E.R., de Queiroz, L.P., Hughes, C.E., 2013. Exploring the tempo of species diversification in legumes. *South African Journal of Botany*, 89:19–30.

Koenen, E.J., Clarkson, J.J., Pennington, T.D. and Chatrou, L.W., 2015. Recently evolved diversity and convergent radiations of rainforest mahoganies (Meliaceae) shed new light on the origins of rainforest hyperdiversity. *New Phytologist*, 207(2):327-339.

Koonin, E.V. and Wolf, Y.I., 2009. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. *Biology Direct*, 4(1):33.

Kursar, T.A., Dexter, K.G., Lokvam, J., Pennington, R.T., Richardson, J.E., Weber, M.G., Murakami, E.T., Drake, C., McGregor, R. and Coley, P.D., 2009. The evolution of antiherbivore defenses and their contribution to species coexistence in the tropical tree genus *Inga*. *Proceedings of the National Academy of Sciences, USA*, 106(43):18073-18078.

Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C.J., Wang, C., Zamani, N. and Grant, B.R., 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539):371.

Larget, B.R., Kotha, S.K., Dewey, C.N. and Ané, C., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22):2910-2911.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21:1095–1109.

Lartillot, N., Brinkmann, H. and Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7(1):S4.

Le, Q., Dang, C., Gascuel, O., 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution*, 29:2921–2936.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R., Martienssen, R.A., Coruzzi, G. & DeSalle, R., 2011. A Functional Phylogenomic View of the Seed Plants. *PLoS Genetetics*, 7: e1002411.

Lemmon, A.R., Emme, S.A. and Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61(5):727-744.

Lemmon, E.M. and Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 44:99-121.

Lewis, G.P., 2005. Legumes of the World. Royal Botanic Gardens Kew.

Li, F.W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J. and Burge, D.O., 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences, USA*, 111(18):6672-6677.

Li, G., Davis, B.W., Eizirik, E. and Murphy, W.J., 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Research*, 26(1):1-11.

Li, H.T., Yi, T.S., Gao, L.M., Ma, P.F., Zhang, T., Yang, J.B., Gitzendanner, M.A., Fritsch:W., Cai, J., Luo, Y. and Wang, H., 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants*, 5(5):461.

Lohaus, R., Van de Peer, Y., 2016. Of dups and dinos: evolution at the K/Pg boundary. *Current Opinions in Plant Biology*, 30:62–69.

LPWG (Legume Phylogeny Working Group), 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon*, 66:44–77.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H. and Burke, J.M., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences*, 2(2):1300085.

Marazzi, B. and Sanderson, M.J., 2010. Large-scale patterns of diversification in the widespread legume genus *Senna* and the evolutionary role of extrafloral nectaries. *Evolution*, 64(12):3570-3592.

Marcet-Houben, M. and Gabaldón, T., 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology*, 13(8):e1002220.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T. and Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, 22(4):746-754.

McKey, D., 1994. Legumes and nitrogen: The evolutionary ecology of a nitrogen-demanding lifestyle. In: Sprent J.I., McKey D., editors. Advances in legume systematics part 5. The nitrogen factor. Richmond, UK: Royal Botanic Gardens, Kew. p. 211–228.

Meier, J.I., Marques, D.A., Mwaiko, S., Wagner, C.E., Excoffier, L. and Seehausen, O., 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, 8:14363.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. and Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541-i548.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G. and Niehuis, O., 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763-767.

Moore, A.J., Vos, J.M.D., Hancock, L.P., Goolsby, E. and Edwards, E.J., 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the portullugo clade (Caryophyllales). *Systematic Biology*, 67(3):367-383.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G. and Worm, B., 2011. How many species are there on Earth and in the ocean? *PLoS Biology*, 9(8):e1001127.

Morales-Briones, D.F., Liston, A. and Tank, D.C., 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytologist*, 218(4):1668-1684.

Mudge, J., Cannon, S.B., Kalo, P., Oldroyd, G.E.D., Roe, B.A., Town, C.D. and Young, N.D., 2005. Highly syntenic regions in the genomes of soybean, *Medicago truncatula*, and *Arabidopsis thaliana*. *BMC Plant Biology*, 5:15.

Nicholls, J.A., Pennington, R.T., Koenen, E.J., Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N. and Kidner, C.A., 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science*, 6:710.

O'Brien, S.J. and Stanyon, R., 1999. Phylogenomics: Ancestral primate viewed. *Nature*, 402(6760):365.

Pagel, M. and Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571-581.

Pease, J.B., Haak, D.C., Hahn, M.W. and Moyle, L.C., 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology*, 14(2):e1002379.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G. and Baurain, D., 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology*, 9(3):e1000602.

Philippe, H., Delsuc, F., Brinkmann, H. and Lartillot, N., 2005. Phylogenomics. *Annual Review of Ecology, Evolution and Systematics*, 36:541-562.

Qiu, Y.L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., Dombrovska, O., Lee, J., Kent, L., Rest, J. and Estabrook, G.F., 2006. The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National Academy of Sciences, USA*, 103(42):15511-15516.

Rabier, C.E., Ta, T. and Ané, C., 2013. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular Biology and Evolution*, 31(3):750-762.

Richardson, J.E., Pennington, R.T., Pennington, T.D. and Hollingsworth, P.M., 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science*, 293(5538):2242-2245.

Rokas, A., Williams, B.L., King, N. and Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798.

Salichos, L. and Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327.

Sass, C., Iles, W.J., Barrett, C.F., Smith, S.Y. and Specht, C.D., 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ*, 4:e1584.

Sayyari, E. and Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, 33(7):1654-1668.

Schrire, B.D., Lewis, G.P., Lavin, M., 2005. Biogeography of the Leguminosae. In: Lewis, G.P., 2005. Legumes of the World. Royal Botanic Gardens Kew. p. 21-54.

Scotland, R.W. and Wortley, A.H., 2003. How many species of seed plants are there? *Taxon*, 52(1):101-104.

Seijo, J.G. and Solis Neffa, V.G., 2004. The cytological origin of the polyads and their significance in the reproductive biology of *Mimosa bimucronata*. *Botanical Journal of the Linnean Society*, 144(3):343-349.

Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M. and Van Heeringen, S.J., 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature*, 538(7625):336.

Smith, S.A., Moore, M.J., Brown, J.W., Yang, Y., 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15:150.

Smith, S.A., Brown, J.W., Walker, J.F., 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS* ONE, 13(5):e0197433.

Sprent, J.I. and Platzmann, J., 2001. Nodulation in legumes (p. 146). Kew: Royal Botanic Gardens.

Spriggs, E.L., Clement, W.L., Sweeney, P.W., Madriñán, S., Edwards, E.J. and Donoghue, M.J., 2015. Temperate radiations and dying embers of a tropical past: the diversification of *Viburnum*. *New Phytologist*, 207(2):340-354.

Stai, J.S., Yadav, A., Sinou, C., Bruneau, A., Doyle, J.J., Fernández-Baca, D. and Cannon, S.B., 2019. *Cercis*: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. *Frontiers in Plant Science*, 10:345.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B. and Durand, D., 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409-i415.

Suh, A., Smeds, L., Ellegren, H., 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology* 13(8):e1002224.

Suh, A., 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zoologica Scripta*, 45:50–62.

Teeling, E.C., Hedges, S.B., 2013. Making the impossible possible: rooting the tree of placental mammals. *Molecular Biology and Evolution*, 30:1999–2000.

Thorne, R.F., 2002. How many species of seed plants are there?. Taxon, 51(3):511-512.

Tiley, G.P., Barker, M.S. and Burleigh, J.G., 2018. Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biology and Evolution*, 10(11):2882-2898.

Tong, K.J., Duchêne, S., Ho, S.Y. and Lo, N., 2015. Comment on “Phylogenomics resolves the timing and pattern of insect evolution”. *Science*, 349(6247):487-487.

Torsvik, V., Sørheim, R. and Goksøyr, J., 1996. Total bacterial diversity in soil and sediment communities—a review. *Journal of Industrial Microbiology*, 17(3-4):170-178.

Valencia, R., Balslev, H. and Miño, G.P.Y., 1994. High tree alpha-diversity in Amazonian Ecuador. *Biodiversity & Conservation*, 3(1):21-28.

van Velzen, R., Doyle, J.J. and Geurts, R., 2018. A Resurrected Scenario: Single Gain and Massive Loss of Nitrogen-Fixing Nodulation. *Trends in Plant Science*.

Vanneste, K., Baele, G., Maere, S., Van de Peer, Y., 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research*, 24(8):1334–1347.

Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A. and Seehausen, O., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, 22(3):787-798.

Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. and Liston, A., 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, 2(9):1400042.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A. and Ruhfel, B.R., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA*, 111(45):E4859-E4868.

Wing, S.L., Herrera, F., Jaramillo, C.A., Gómez-Navarro, C., Wilf, P. and Labandeira, C.C., 2009. Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest. *Proceedings of the National Academy of Sciences, USA*, 106(44):18627-18632.

Wolfe, K.H. and Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708.

Yang, Z. and Rannala, B., 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303.

Yang, Y., Moore, M.J., Brockington, S.F., Soltis, D.E., Wong, G.K.S., Carpenter, E.J., Zhang, Y., Chen, L., Yan, Z., Xie, Y. and Sage, R.F., 2015. Dissecting molecular evolution in

the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution*, 32(8):2001-2014.

Yang, Y., Moore, M.J., Brockington, S.F., Mikenas, J., Olivieri, J., Walker, J.F., Smith, S.A., 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events within Caryophyllales, including two allopolyploidy events. *New Phytologist*, 217:855–870.

Zachos, J., Pagani, M., Sloan, L., Thomas, E. and Billups, K., 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, 292(5517):686-693.

Zwaenepoel, A. and Van de Peer, Y., 2019. Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates. *Molecular Biology and Evolution*.

Chapter I

LARGE-SCALE GENOMIC SEQUENCE DATA SUPPORT A NEAR-SIMULTANEOUS EVOLUTIONARY ORIGIN OF ALL SIX LEGUME SUBFAMILIES

Authors:
Erik J.M. Koenen[1], Dario I. Ojeda[2,3], Royce Steeves[4,5], Jérémy Migliore[2], Freek T. Bakker[6], Jan J. Wieringa[6,7], Catherine Kidner[8,9], Olivier J. Hardy[2], R. Toby Pennington[8,10], Anne Bruneau[4] and Colin E. Hughes[1]

[1] Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstrasse 107, CH-8008, Zurich, Switzerland
[2] Service Évolution Biologique et Écologie, Faculté des Sciences, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050, Brussels, Belgium
[3] Norwegian Institute of Bioeconomy Research, Høgskoleveien 8, 1433 Ås, Norway
[4] Institut de Recherche en Biologie Végétale and Département de Sciences Biologiques, Université de Montréal, 4101 Sherbrooke St E, Montreal, QC H1X 2B2, Canada
[5] Fisheries & Oceans Canada, Gulf Fisheries Center, 343 Université Ave, Moncton, NB E1C 5K4, Canada
[6] Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
[7] Naturalis Biodiversity Center, Leiden, Darwinweg 2, 2333 CR, Leiden, The Netherlands
[8] Royal Botanic Gardens Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, U.K.
[9] School of Biological Sciences, University of Edinburgh, King's Buildings, Mayfield Rd, Edinburgh, UK
[10] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX4 4RJ, U.K.

Summary

- Phylogenomics is increasingly used to infer deep-branching relationships while revealing the complexity of evolutionary processes such as incomplete lineage sorting, hybridization/introgression and polyploidization. We investigate the deep-branching relationships among subfamilies of the Leguminosae (or Fabaceae), the third largest angiosperm family. Despite their ecological and economic importance, a robust phylogenetic framework for legumes based on genome-scale sequence data is lacking.
- We generated alignments of 72 chloroplast genes and 7,621 homologous nuclear encoded proteins, for 157 and 76 taxa, respectively. We analysed these with Maximum Likelihood, Bayesian Inference, and a multi-species coalescent summary method, and evaluated support for alternative topologies across gene trees.
- The earliest dichotomies in the legume phylogeny are difficult to resolve due to lack of phylogenetic signal across chloroplast genes and the majority of nuclear genes. Strongly supported conflict in the remainder of nuclear genes is suggestive of incomplete lineage sorting or introgression.
- All six subfamilies originated nearly simultaneously, suggesting that the prevailing view of some subfamilies as “basal” or “early-diverging” with respect to others should be abandoned, which has important implications for understanding the evolution of legume diversity and traits. Our study highlights the limits to phylogenetic resolution in relation to rapid successive speciation.

Keywords: Fabaceae, gene tree conflict, incomplete lineage sorting, lack of phylogenetic signal, Leguminosae, phylogenomics

Introduction

Phylogenomic studies often focus on difficult to resolve, deep relationships in the Tree of Life, e.g. in the land plants (Wickett et al., 2014), the deep-branching relationships of animals (Simion et al., 2017), the root of Placentalia (Morgan et al., 2013; Romiguier et al., 2013) and the initial radiation of Neoaves (Suh, 2016). What these studies have shown is that many of these relationships remain unresolved even when deploying large genome-scale DNA

sequence data sets, owing to lack of phylogenetic signal and/or conflicting signals between different genomic regions (Rokas et al., 2003; Salichos & Rokas, 2013), such that the inferred relationships are often only implied by a small fraction of genes or characters (Shen et al., 2017). While fully resolved phylogenies will therefore likely remain elusive, phylogenomic analysis can provide important insights into the evolutionary processes that shape phylogeny and the underlying causes of lack of phylogenetic resolution. For instance, incomplete lineage sorting (ILS) or deep coalescence is widely recognised as a process causing phylogenetic discordance among gene trees and is routinely invoked to explain conflicting genealogies, even though few studies have provided compelling evidence for it (Suh et al., 2015). Lack of phylogenetic signal and gene tree estimation errors may be equally or more important (Scornavacca & Galtier, 2017). Hybridization and/or whole genome duplication (WGD) are other processes that can complicate phylogenetic analyses because these phenomena violate the assumption of a bifurcating topology and/or create difficulties for the inference of homology of molecular characters due to paralogy. While introgression through hybridization leads to reticulate patterns (a “network”), both a lack of phylogenetic signal and gene tree conflict due to ILS can cause hard polytomies. It can be difficult to determine whether a polytomy should be viewed as ‘soft’ in the case of insufficient data, or ‘hard’ in the case of (nearly) simultaneous speciation (Suh, 2016), since the latter is often implied by a mere absence of evidence for resolved relationships, rather than convincing evidence in favour of simultaneous speciation. Especially for deep divergences, polytomies and reticulate patterns are expected to be difficult to analyse due to the erosion of phylogenetic signal over time by saturation of substitutions.

The legume family (Leguminosae or Fabaceae) is one of the most prominent angiosperm families across global ecosystems and with c. 20,000 spp. (Lewis et al., 2005) it ranks third in size after the orchids (Orchidaceae) and daisies (Asteraceae). More than three decades of phylogenetic research since the first molecular phylogenies of the family were inferred (Doyle, 1995; Doyle et al., 1997; reviewed in LPWG, 2013a) has culminated in the recent reclassification of the Leguminosae into six subfamilies, with diverse floral morphologies (Fig. 1a-f; LPWG, 2017). The defining feature of the family is the typical unicarpellate and unilocular superior fruit, which is referred to as the “legume” or “pod” (Fig.

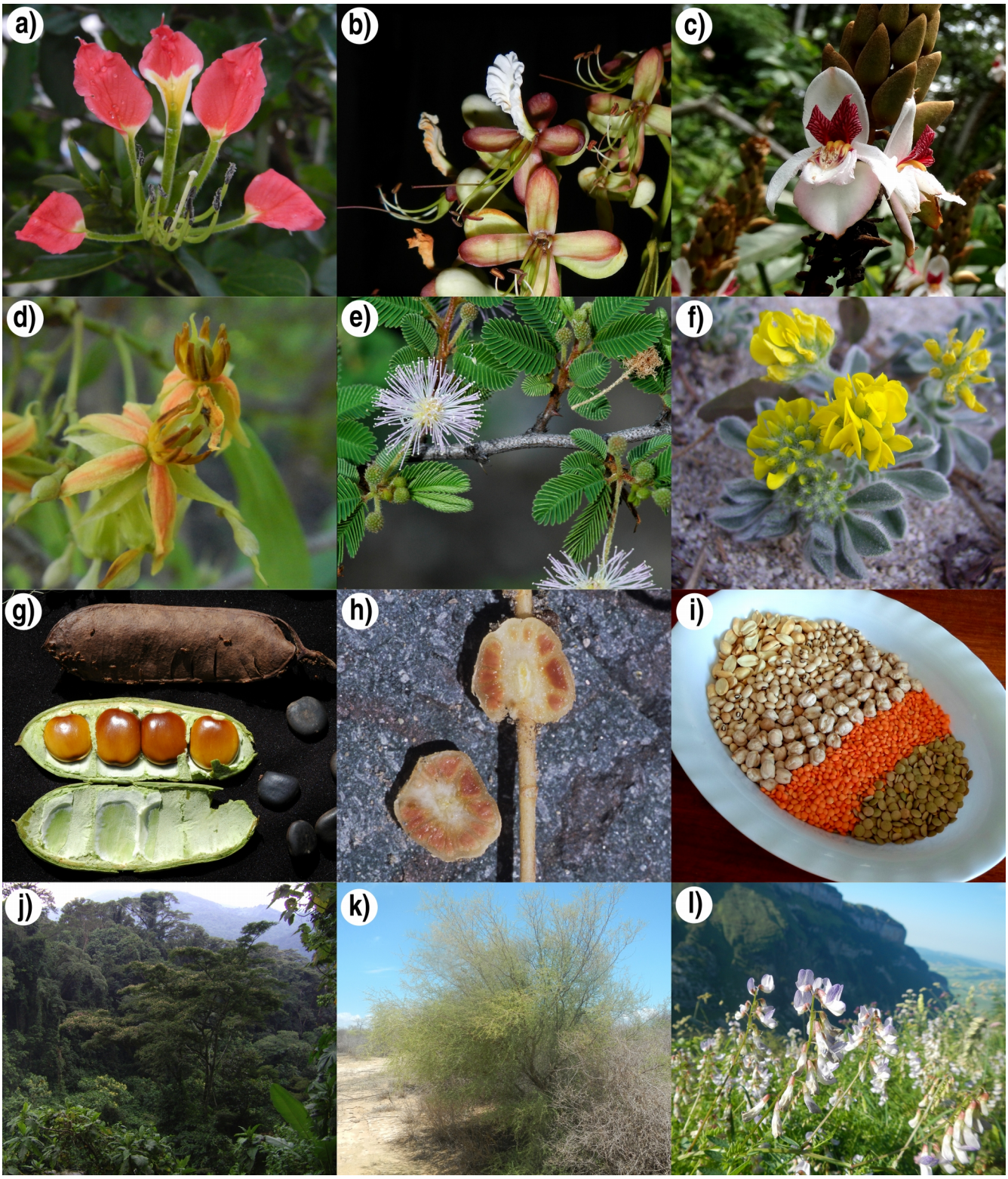


Figure 1. Diversity, ecology and economic importance of legumes. The family is subdivided into subfamilies (a) Cercidoideae (*Bauhinia madagascariensis*), (b) Detarioideae (*Macrolobium* sp.), (c) Duparquetioideae

(*Duparquetia orchidacea*), (d) Dialioideae (*Baudouinia* sp.), (e) Caesalpinioideae (*Mimosa pectinatipinna*) and (f) Papilionoideae (*Medicago marina*). While the family has a very diverse floral morphology, the fruit (g) (*Brodriguesia santosii*), which comes in many shapes and is most often referred to as 'pod' or 'legume', is the defining feature of the family. A large fraction of legume species is known to fix atmospheric nitrogen symbiotically with 'rhizobia', bacteria that are incorporated in root nodules, for example in *Lupinus nubigenus* (h). Economically, the family is the second most important of flowering plants after the grasses, with a wide array of uses, including timber, ornamentals, fodder crops, and notably, pulse crops such as peanuts (Arachis), beans (Phaseolus), chickpeas (Cicer) and lentils (Lens) (i). Ecologically, legumes are also extremely diverse and important, occurring and often dominating globally across disparate ecosystems, including wet tropical forest, for example *Albizia grandibracteata* in the East African Albertine Rift (j), savannas, seasonally dry tropical forests, and semi-arid thorn-scrub, for example *Mimosa delicatula* in Madagascar (k) and temperate woodlands and grasslands, for example *Vicia sylvatica* in the European Alps (l). -- Photos a, b, d, f, i, j, k, l by Erik Koenen, c by Jan Wieringa and e, g, h by Colin Hughes.

1g), a character shared across all subfamilies. Legumes are the second most cultivated plant family after the Poaceae, and its species serve many purposes for humans, including timber, ornamentals, fodder crops, hallucinogens, medicines, and most notably, a large set of globally important pulse crops (Fig. 1i). A key trait of many legumes is the ability to fix atmospheric nitrogen via symbiosis with “rhizobia”-bacteria in root nodules (Fig. 1h), which leads to enriched soils, high nitrogen content in the leaves, and protein-rich seeds (McKey, 1994). Furthermore, legume species are omnipresent and often abundant in nearly all vegetation types across the planet, ranging in habit from large rainforest trees to small temperate herbs, representing one of the most spectacular examples of evolutionary and ecological radiation of any angiosperm family (Fig. 1j-l).

Despite this prominence, a well-resolved phylogenetic framework for the family, based on genome-scale data, is lacking and the origin and early evolution, including deep-branching relationships among the six legume subfamilies are poorly understood, hampering research in comparative legume biology. Sister-group relationships between subfamilies Papilionoideae and Caesalpinioideae (sensu LPWG, 2017), and of the clade combining these two subfamilies with the newly recognized subfamily Dialioideae, have been recovered previously (Lavin et al., 2005; Bruneau et al., 2008; LPWG, 2017). However, the relationships between the clade comprising those three subfamilies and the other three subfamilies Cercidoideae,

Detarioideae and Duparquetioideae have not been resolved with high confidence (cf. Bruneau et al., 2008; LPWG, 2017). Moreover, previous legume phylogenies have been exclusively inferred from a handful of chloroplast markers (Doyle et al., 1997; Wojciechowski et al., 2004; Lavin et al., 2005; Bruneau et al., 2008; Simon et al., 2009; Cardoso et al., 2012, 2013; LPWG, 2017), even though it is preferable to infer species trees based on analysis of unlinked nuclear loci to account for different evolutionary histories of individual genes (Maddison, 1997).

Alongside improving resolution in the legume phylogeny, our main objective is to investigate the causes of the lack of resolution surrounding the initial divergence and deep-branching relationships of legumes. Specifically we ask whether lack of phylogenetic signal or conflicting evolutionary relationships across different genomic elements is causing lack of resolution, or whether previous studies simply did not analyse sufficiently large data sets. Following on from this, given a sufficiently large volume of data, can we reject a hard polytomy and find support for a fully bifurcating topology? In addition to analysing sequences of nearly all protein-coding genes from the chloroplast genome, we analyse thousands of gene alignments from the nuclear genome, with a total aligned length that is several orders of magnitude longer than those previously used in legume phylogenetics. This means we can also dissect and analyse phylogenetic signal and conflict across unlinked loci, and data deficiency can most likely be ruled out as the cause of lack of resolution. We analyse these new datasets with Maximum Likelihood (ML) analysis, Bayesian inference and a multi-species coalescent summary method to infer the most likely relationships among the legume subfamilies. Having inferred the most likely species-tree topology, we evaluate numbers of supporting and conflicting bipartitions across gene trees for critical nodes, and discuss the implications for our understanding of the early evolution of legumes.

Materials & Methods

DNA/RNA extraction and sequencing

For the newly generated chloroplast gene data, DNA was extracted from fresh leaves, leaf tissue preserved in silica-gel or herbarium specimens, using the Qiagen DNeasy Plant Mini Kit. Sequencing libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina, and sequenced on the Illumina HiSeq 2000 sequencing platform, at low coverage ('genome-skimming'), or as part of hybrid capture experiments for a separate study on mimosoid legumes (Koenen et al., unpublished data). For the newly generated nuclear gene data, we used transcriptome sequencing, for which RNA was extracted from fresh leaves using the Qiagen RNeasy Plant Mini Kit. RNA sequencing libraries were prepared using the Illumina TruSeq RNA Library Prep Kit and sequenced on the Illumina HiSeq 2000 sequencing platform. All lab procedures were performed according to the specifications and protocols provided by manufacturers of the kits.

Sequence assembly

Raw reads for the chloroplast DNA data were cleaned and filtered using the following steps: (1) Illumina adapter sequence artifacts were trimmed using Trimmomatic v. 0.32 (Bolger et al., 2014), (2) overlapping read pairs were merged with PEAR v. 0.9.8 (Zhang et al., 2014) and (3) low quality reads were discarded and low quality bases at the end of the reads were trimmed with Trimmomatic v. 0.32 (using settings MAXINFO:40:0.1 LEADING:20 TRAILING:20). The quality-filtered reads were assembled into contigs using the SPAdes assembler v. 3.6.2 (Bankevich et al., 2012). For RNA data, we used the FASTX-toolkit v. 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to remove low quality reads (less than 80% of bases with a quality score of 20 or higher), TagDust v. 1.12 (Lassmann et al., 2009) to remove adapter sequences, and PRINSEQ-lite v. 0.20.4 (Schmieder & Edwards, 2011) to trim low quality bases off the ends of reads. Transcriptome assembly was performed

on the quality-filtered reads using Trinity (Grabherr et al., 2011; Release 2012-06-08), with default settings.

Chloroplast proteome alignment

DNA sequences of protein-coding chloroplast genes were newly generated for 49 accessions, or extracted from several different data sources, as specified in Table S1. Sequence data were extracted directly from annotated plastomes in Genbank, by blast searches from *de novo* assembled contigs and from transcriptomes using custom Python scripts. Sequences for some outgroup taxa (data from Moore et al., 2010) were downloaded separately per gene from Genbank. For each gene, a codon alignment was inferred using MACSE v. 1.01b (Ranwez et al., 2011). Phylogenetic trees were inferred for each gene separately to screen for erroneously aligned sequences with RAxML v. 8.2 (Stamatakis, 2014). For some species, individual gene sequences that led to anomalously long terminal branches were removed. The genes *accD* and *clpP* were removed completely, because they have been lost, pseudogenized or relocated to the nuclear genome in several legume lineages (Magee et al., 2010; Williams et al., 2015; Dugas et al., 2015), leading to poor quality alignments. Gene alignments were concatenated, the full alignment visually checked, and obvious misalignments resolved. Furthermore, sequence errors (single A/T indels) that caused frameshift mutations were corrected and the accuracy of the alignment at codon level was assessed and corrected if necessary. For the genes *ndhF*, *ndhI*, *rpl20* and *rps18*, where the ends of coding sequences had varying lengths, all sites between the first and last stop codon in the alignment were excluded, since they were poorly aligned. Finally, using BMGE v. 1.12 (Block Mapping and Gathering with Entropy; Criscuolo & Gribaldo, 2010) the codon alignment was translated to amino acid sequences.

Nuclear gene data and matrix assembly

Whole genome and transcriptome data were downloaded from various sources and augmented with newly generated transcriptome sequence data for six Caesalpinioideae and

Detarioideae taxa (Table S2). Peptide sequences were downloaded from annotated genomes or were extracted from transcriptome assemblies using TransDecoder (http://transdecoder.github.io/). To assemble the nuclear peptide sequence data into aligned gene matrices, we used the pipeline of Yang & Smith (2014). We performed mcl clustering as described in Yang & Smith (2014), with a hit fraction cut-off of 0.75, inflation value of 2 and a minimum log-transformed e-value of 30. These settings lead to clusters with good overlap between sequences and good alignability (omitting genes that are too variable), although we may have lost a few short gene clusters. The resulting homolog gene clusters were subjected to two rounds of alignment with MAFFT v. 7.187 (Katoh & Standley, 2013), gene tree inference with RaxML v. 8.2 (Stamatakis, 2014), pruning and masking of tips and cutting deep paralogs as described in Yang & Smith (2014). In the first round we used 0.3 and 1.0 as relative and absolute cut-offs for trimming tips, respectively, and 0.5 as the minimum cut-off for cutting deep paralogs, and keeping all clusters with a minimum of 25 taxa for the second round. In the second round we used more stringent cut-off values (0.2 and 0.5 for trimming tips and 0.4 for cutting deep paralogs). See Yang & Smith (2014) for more information on these parameter settings. One-to-one orthologs, i.e. those homolog gene clusters in which each taxon is represented only by a single gene copy, and rooted ingroup (RT) homologs were extracted from the homolog cluster trees, with a minimum aligned length of 100 amino acids for each homolog. RT homologs are extracted by orienting homolog cluster trees by rooting them on the outgroup (in our case *Aquilegia coerulea* and *Papaver somniferum*), detecting gene duplications and pruning the paralog copies with fewer taxa present until each taxon is represented by a single copy. The outgroup is also pruned, and clusters in which each taxon is only present once are also included, meaning that all 1-to-1 orthologs are also in the RT homolog set. See Yang & Smith (2014) for a more details. Sequences with more than 50% gaps and all sites with more than 5% missing data were removed from the homolog alignments using BMGE. For the 1-to-1 orthologs, alignments with fewer than 50 taxa were discarded. For the larger set of RT homologs, alignments with fewer than 25 taxa were discarded.

Phylogenetic inferences

Gene tree inferences were made with Maximum Likelihood (ML) analysis in RaxML v. 8.2 (Stamatakis, 2014). Species tree analyses were performed with ML in RAxML, using Bayesian inference in Phylobayes-MPI 1.7 (Lartillot et al., 2013) and the multi-species coalescent summary method implemented in ASTRAL v. 5.6.3 (Mirarab et al., 2014).

Gene trees of 1-to-1 orthologs and RT homologs were estimated with RAxML using the WAG + G model, with 100 rapid bootstrap replicates. We calculated 80% majority-rule consensus trees for each ortholog or homolog and used these to calculate Internode Certainty All (ICA) values using RAxML, in order to include only nodes that received 80% or greater bootstrap support (BS) in the individual gene trees. We also used the concordance analysis in phyparts (Smith et al., 2015), with a BS cut-off of 50% and used the output to extract numbers of supporting and conflicting bipartitions for plotting pie charts on the species tree.

For the ML species tree analysis using nucleotide sequences of the chloroplast alignment, we used PartitionFinder 2 (Lanfear et al., 2017) to estimate partitions, with a minimum length per partition set to 500 nucleotides, and allowing different codon positions per gene to be in different partitions. The resulting 16 partitions were run with the GTR + GAMMA model, and 1000 rapid bootstrap replicates were carried out. For the amino acid sequences, the ML analyses of both the chloroplast alignment and the concatenated alignment of nuclear 1-to-1 orthologs were analysed with the LG4X model, without partitioning, as this model accounts for substitution rate heterogeneity across the alignment by estimating four different LG substitution matrices (Le et al., 2012). For the chloroplast alignment, 1000 rapid bootstrap replicates were carried out.

Bayesian species tree analyses were performed with the CATGTR model, with invariant sites deleted and default settings for other options in Phylobayes. Analyses on the chloroplast alignment were run until the chain reached convergence (usually after 10-20k cycles), and the first 10% of the chain was discarded as burnin. To perform Bayesian analyses on the complete 1-to-1 ortholog set in a computationally tractable manner, we ran 25 gene jack-knifing replicates without replacement, dividing the total number of genes over five subsets with five replicates. These subsampled replicates were run in Phylobayes-MPI, with a

starting tree derived from the analysis sampling the 100 genes with the longest gene tree length, using the CATGTR model with constant sites deleted, for 1000 cycles each. We found that all 25 chains had converged after a few hundred cycles, and discarded the first 500 cycles of each as burn-in. A majority-rule consensus tree was constructed using sumtrees.py (from the Dendropy library; Sukumaran & Holder, 2010) from 12500 total posterior trees, representing the MCMC cycles 501-1000 of each replicate. For the ML and Bayesian analyses, concatenated alignments were not partitioned. Instead we rely on the LG4X and CATGTR models to take rate heterogeneity into account, since these models describe heterogeneity across alignments more accurately than partitioning by gene and/or codon as the substitution process also varies across gene sequences and codon positions.

For the multi-species coalescent analysis with ASTRAL, we used the 1-to-1 ortholog gene trees estimated with RAxML, using local posterior probability and quartet support to evaluate the inferred topology (Sayyari & Mirarab, 2016). We also used the polytomy test in ASTRAL (Sayyari & Mirarab, 2018) to evaluate whether a hard polytomy can be rejected for the relationships among subfamilies, where a *p*-value of <0.05 is considered as a rejection of the null hypothesis of a polytomy.

We used SplitsTree4 to draw a filtered supernetwork (Whitfield et al., 2008) of the 1,103 1-to-1 orthologs, using the 80% majority-rule consensus trees to only include well-supported bipartitions to infer the network. All gene trees were pruned to include only the nitrogen-fixing clade of angiosperms. Furthermore, for the relatively densely sampled Papilionoideae and Caesalpinioideae we several taxa that were less well represented across gene trees. The mintrees parameter was set to 552 (at least 50% of the number of orthologs) and the maximum distortion parameter was set to 0.

Counting supporting bipartitions for key nodes across gene trees

Using a custom python script (Supplementary Script S1), numbers of matching and alternative bipartitions across the RT gene trees were counted for particular nodes labelled A-H in Figure 3 in the legume phylogeny. For this purpose, we assessed monophyly of each of the subfamilies and combinations (clades) of subfamilies, against the outgroup, across all

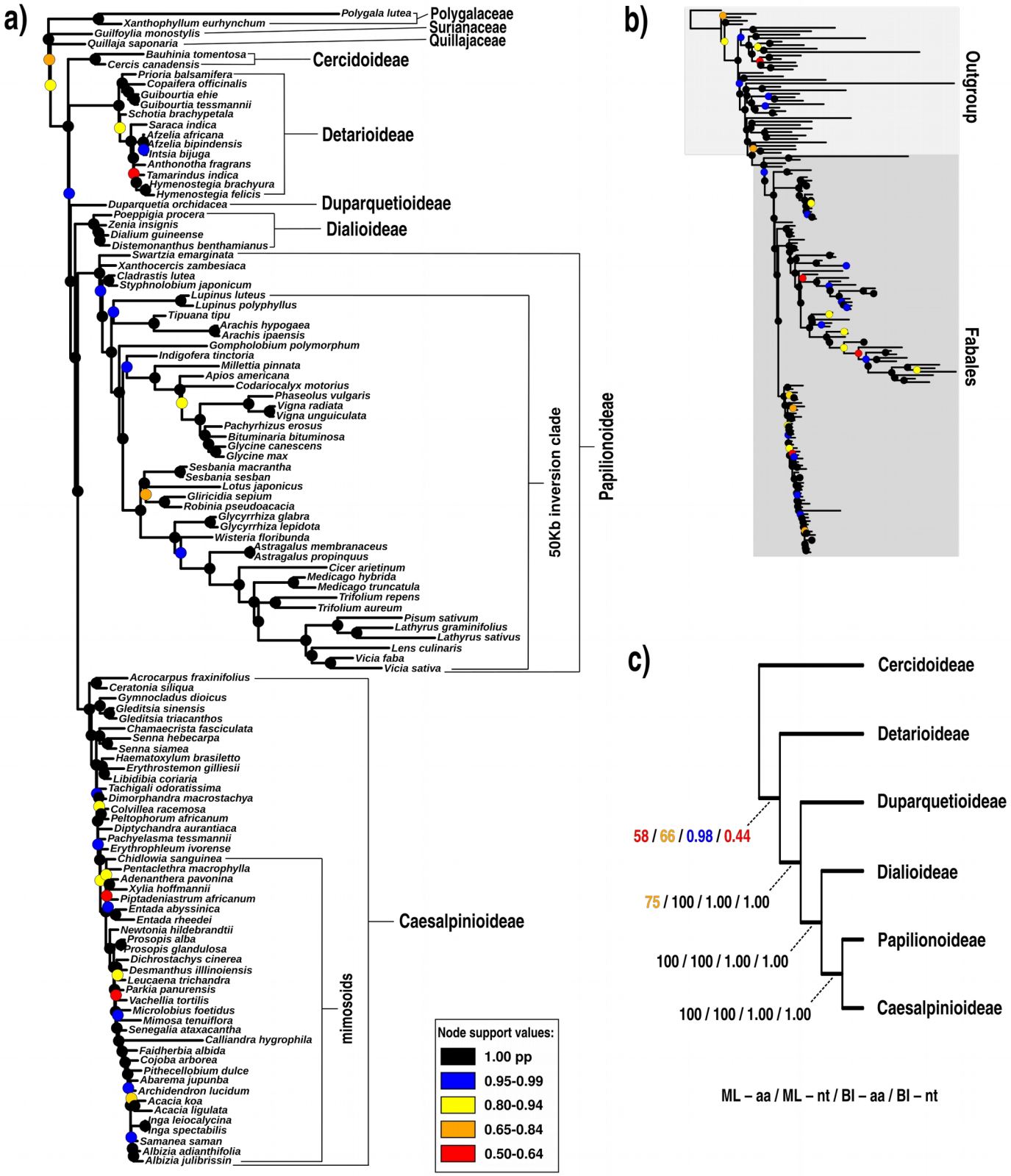
gene trees. For each gene tree, we first assessed whether all six groups (five subfamilies plus the outgroup) are represented, gene trees with missing groups were not taken into account. Next, we evaluated whether the gene tree includes a matching bipartition for the family, each subfamily and for all possible combinations of subfamilies. A matching bipartition means that all taxa of a subfamily or combination of subfamilies are separated from all other taxa in the gene tree, including the outgroup, thus constituting support for that clade to be monophyletic. For combinations of subfamilies, the subfamilies themselves do not necessarily need to be monophyletic, but all taxa within the combined subfamilies should be separated from all other taxa to constitute a matching bipartition, and thus to be a supported clade in the gene tree. For a clade to be well supported, we expect matching bipartitions for a majority of gene trees. For clades to be poorly supported, we expect gene trees to be either uninformative due to low phylogenetic signal or to contain significant conflicting bipartitions, hence relatively low numbers of matching bipartitions. All counts were done for ML gene trees of RT homologs, and with 50 and 80% bootstrap cut-offs.

Results

The chloroplast alignment includes 72 protein-coding genes, for 157 taxa (including 111 legume species; Table S1), with a total aligned length of 75,282 bp or 25,094 amino acid residues. From transcriptomes and fully sequenced genomes, we gathered 9,282 homologous nuclear encoded gene clusters for 76 taxa including 42 legume species (Table S2). From these clusters, we extracted protein alignments of 1,103 1-to-1 orthologs for species tree inference with a total aligned length of 325,134 amino acids when concatenated, and 7,621 Rooted Ingroup (RT) homologs for additional gene tree inference. The alignments, gene trees and species trees are available in TreeBASE (accession number XXXX) and on Dryad (Supplementary Data Files S1-10; doi: XXXX).

Inferring the species tree

Our analyses reveal that both the chloroplast and nuclear data sets resolve all subfamilies as monophyletic with full support. Most relationships among the subfamilies are


a)
Polygala lutea
Xanthophyllum eurhynchum
Guilfoylia monostylis
Quillaja saponaria
Polygalaceae
Surianaceae
Quillajaceae
Bauhinia tomentosa
Cercis canadensis
Cercidoideae
Prioria balsamifera
Copaifera officinalis
Guibourtia ehie
Guibourtia tessmannii
Schotia brachypetala
Saraca indica
Afzelia africana
Afzelia bipindensis
Intsia bijuga
Anthonotha fragrans
Tamarindus indica
Hymenostegia brachyura
Hymenostegia felicis
Detarioideae
Duparquetia orchidacea
Duparquetioideae
Poeppigia procera
Zenia insignis
Dialium guineense
Distemonanthus benthamianus
Dialioideae
Swartzia emarginata
Xanthocercis zambesiaca
Cladrastis lutea
Styphnolobium japonicum
Lupinus luteus
Lupinus polyphyllus
Tipuana tipu
Arachis hypogaea
Arachis ipaensis
Gompholobium polymorphum
Indigofera tinctoria
Millettia pinnata
Apios americana
Codariocalyx motorius
Phaseolus vulgaris
Vigna radiata
Vigna unguiculata
Pachyrhizus erosus
Bituminaria bituminosa
Glycine canescens
Glycine max
Sesbania macrantha
Sesbania sesban
Lotus japonicus
Gliricidia sepium
Robinia pseudoacacia
Glycyrrhiza glabra
Glycyrrhiza lepidota
Wisteria floribunda
Astragalus membranaceus
Astragalus propinquus
Cicer arietinum
Medicago hybrida
Medicago truncatula
Trifolium repens
Trifolium aureum
Pisum sativum
Lathyrus graminifolius
Lathyrus sativus
Lens culinaris
Vicia faba
Vicia sativa
50Kb inversion clade
Papilionoideae
Acrocarpus fraxinifolius
Ceratonia siliqua
Gymnocladus dioicus
Gleditsia sinensis
Gleditsia triacanthos
Chamaecrista fasciculata
Senna hebecarpa
Senna siamea
Haematoxylum brasiletto
Erythrostemon gilliesii
Libidibia coriaria
Tachigali odoratissima
Dimorphandra macrostachya
Colvillea racemosa
Peltophorum africanum
Diptychandra aurantiaca
Pachyelasma tessmannii
Erythrophleum ivorense
Chidlowia sanguinea
Pentaclethra macrophylla
Adenanthera pavonina
Xylia hoffmannii
Piptadeniastrum africanum
Entada abyssinica
Entada rheedei
Newtonia hildebrandtii
Prosopis alba
Prosopis glandulosa
Dichrostachys cinerea
Desmanthus illinoiensis
Leucaena trichandra
Parkia panurensis
Vachellia tortilis
Microlobius foetidus
Mimosa tenuiflora
Senegalia ataxacantha
Calliandra hygrophila
Faidherbia albida
Cojoba arborea
Pithecellobium dulce
Abarema jupunba
Archidendron lucidum
Acacia koa
Acacia ligulata
Inga leiocalycina
Inga spectabilis
Samanea saman
Albizia adianthifolia
Albizia julibrissin
mimosoids
Caesalpinioideae
Node support values:
1.00 pp
0.95-0.99
0.80-0.94
0.65-0.84
0.50-0.64
b)
Outgroup
Fabales
c)
Cercidoideae
Detarioideae
Duparquetioideae
Dialioideae
Papilionoideae
Caesalpinioideae
58 / 66 / 0.98 / 0.44
75 / 100 / 1.00 / 1.00
100 / 100 / 1.00 / 1.00
100 / 100 / 1.00 / 1.00
ML – aa / ML – nt / BI – aa / BI – nt

(previous page) Figure 2. Phylogeny of legumes based on Bayesian analyses of 72 protein coding chloroplast genes under the CATGTR model in Phylobayes. (a) Majority-rule consensus tree of the amino acid alignment, showing only the Fabales portion of the tree, outgroup taxa pruned, (b) complete tree including outgroup taxa, (c) simplified tree showing support for subfamily relationships with different inference methods (ML = Maximum Likelihood, BI = Bayesian Inference) and sequence types (aa = amino acids, nt = nucleotides). Majority-rule consensus trees for both the amino acid and nucleotide alignments with tip labels for all taxa and support values indicated are included in supporting information (Figs S1-2).

also robustly resolved (Figs 2, 3, 4 & S1-7), with the notable exception of the root node. The clade consisting of Papilionoideae, Caesalpinioideae and Dialioideae is recovered in all analyses, with *Duparquetia* as the sister-group to this clade as inferred from chloroplast data. *Duparquetia* is not sampled for nuclear data, therefore transcriptome or genome sequencing is necessary for this taxon to confirm the relationship found by chloroplast data. The root node of the legume family is more difficult to resolve, and the chloroplast and nuclear data sets estimate conflicting topologies. The chloroplast alignment weakly supports Cercidoideae as sister to the rest of the family (Figs 2c & S1-S4), except the Bayesian analysis of nucleotide sequences. To resolve deep divergences, amino acid sequences are more suitable because they are less saturated with substitutions (silent substitutions are absent), and therefore less prone to long branch attraction (LBA). Additionally, the LG4X and CAT models better account for heterogeneous substitution rates across sites in the alignment (Lartillot & Philippe, 2004; Le et al., 2012). Taken together, this suggests that the sister-group relationship of Cercidoideae with the rest of the family is the most likely rooting as inferred from chloroplast data, but given the low BS values and lack of resolution in the Bayesian analysis of nucleotide sequences, phylogenetic signal in the chloroplast data with regards to the root node appears to be limited.

In contrast to the chloroplast phylogeny, in all analyses of the 1,103 nuclear 1-to-1 orthologs, we recover a sister-group relationship between Cercidoideae and Detarioideae, with this clade sister to the clade comprising Dialioideae, Caesalpinioideae and Papilionoideae (note that Duparquetioideae is not sampled) (Figs 3, 4 & S5-7). We inferred a ML tree of the concatenated alignment with the LG4X model, and calculated Internode Certainty All (ICA) values from bootstrapped gene trees on this topology (Fig. 3 & S5), for

which only gene tree bipartitions that received at least 80% BS were considered. The internode certainty metric was introduced to assess phylogenetic conflict among loci and identify internodes with high certainty, particularly in phylogenomic studies where bootstrap values are often inflated (Salichos & Rokas, 2013). The sister-group relationship between Cercidoideae and Detarioideae is well-supported, receiving an ICA value of 0.85. A Bayesian jackknifing analysis with the CATGTR model infers a nearly identical topology to the ML topology (Fig. 4a & S6), with posterior probability of 0.91 in support of this same relationship. The multi-species coalescent species-tree inferred with ASTRAL (Mirarab et al., 2014), which accounts for incomplete lineage sorting (ILS), is also consistent with that relationship (Fig. 4b & S7), with the Cercidoideae/Detarioideae clade supported by a local posterior probability of 0.95 (Sayyari & Mirarab, 2016). In summary, all analyses of nuclear protein alignments lend strong support for a sister-group relationship between Cercidoideae and Detarioideae.

While not the primary focus of this study, it is worth noting that the deep-branching relationships inferred within subfamilies are mostly congruent with those found in previous analyses based on only one or two chloroplast markers (Bruneau et al. 2008; Cardoso et al. 2012, 2013; LPWG 2017), and although taxon sampling remains sparse, our data add further resolution (Figs 2,3 & S1-7). The chloroplast matrix has better sampling within subfamilies and a few notable relationships are inferred (Suppl. Figs S1-S4): the Cassia and Caesalpinia clades are not sister clades as recovered by Bruneau et al. (2008) and Manzanilla & Bruneau (2012), but instead form successive sister-groups to the clade comprising mimosoids and genera of the Dimorphandra group and the Peltophorum and Tachigali groups; the Swartzieae (represented here by *Swartzia emarginata*) are sister to the rest of the Papilionoideae as was also recovered by Pennington et al. (2001), not the ADA clade (represented here by *Xanthocercis zambesiaca*) as inferred (albeit with low support) by Cardoso et al. (2013). Also, a novel sister-group relationship between genistoids s.l. and dalbergioids s.l. is found in the chloroplast phylogeny and concatenated nuclear analyses (Suppl. Figs S1-S6), a relationship not present in previous analyses (Lavin et al., 2005; Cardoso et al., 2013). These results demonstrate the importance of evaluating previously inferred legume relationships with genome-scale data, something that is already being addressed in phylogenomic studies of several of the five non-monotypic subfamilies.

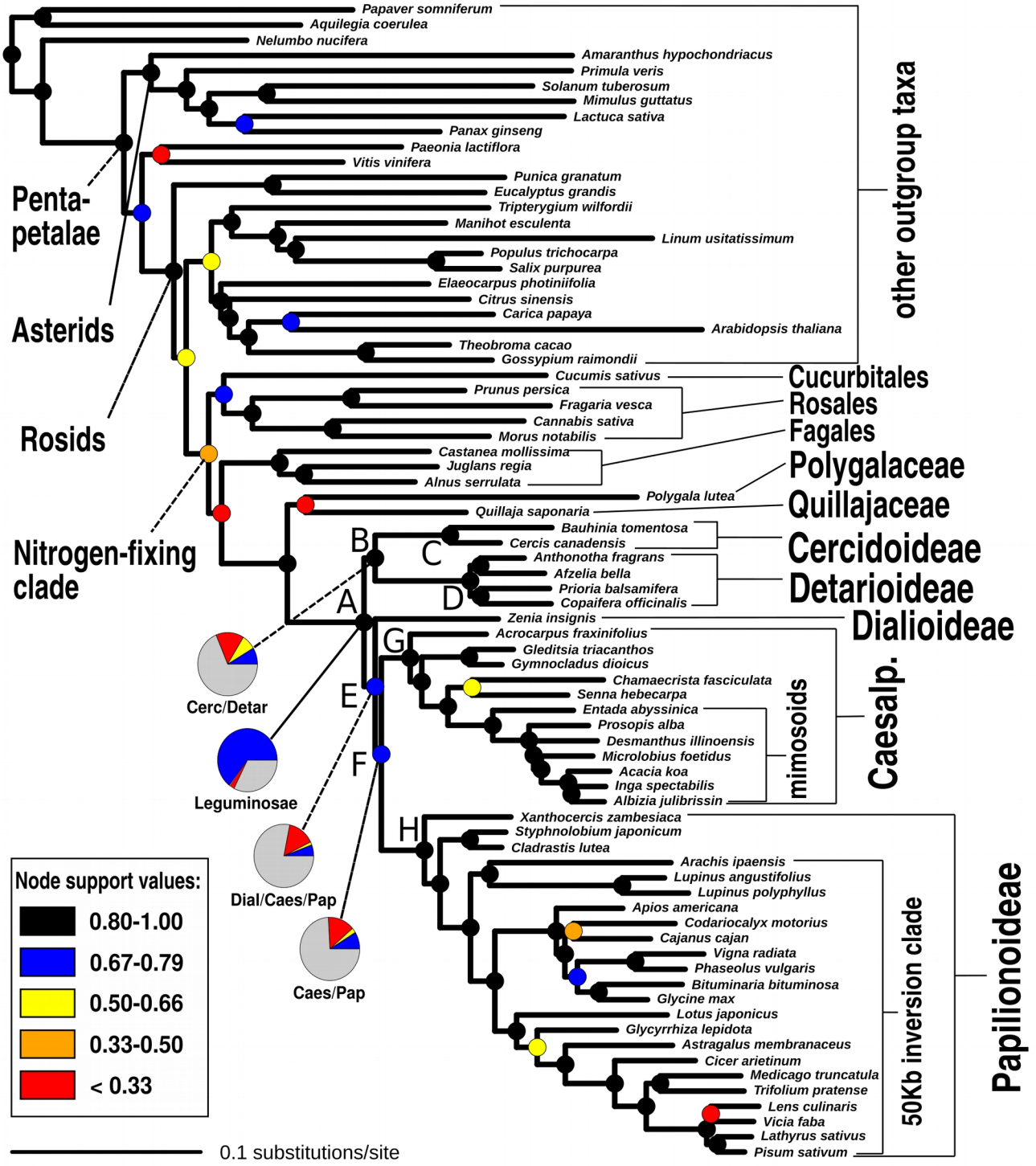


Figure 3. Maximum Likelihood phylogeny of legumes estimated with RAxML under the LG4X model from a concatenated alignment of 1,103 nuclear orthologs. Internode Certainty All (ICA) values are indicated with

coloured symbols on nodes for simplicity of presentation, see Figure S5 for actual support values for all nodes. For the first four divergences in the legume family, pie charts indicate the proportions of gene trees supporting the relationship shown (blue), supporting the most prevalent conflicting bipartition (yellow), supporting other conflicting bipartitions (red) and uninformative genes (i.e. no bootstrap support (BS) and/or missing relevant taxa; grey). Numbers of bipartitions for the pie charts are derived from Phyparts analyses with a 50% BS filter. Labelled nodes A-H are analysed in more detail in Figure 6.

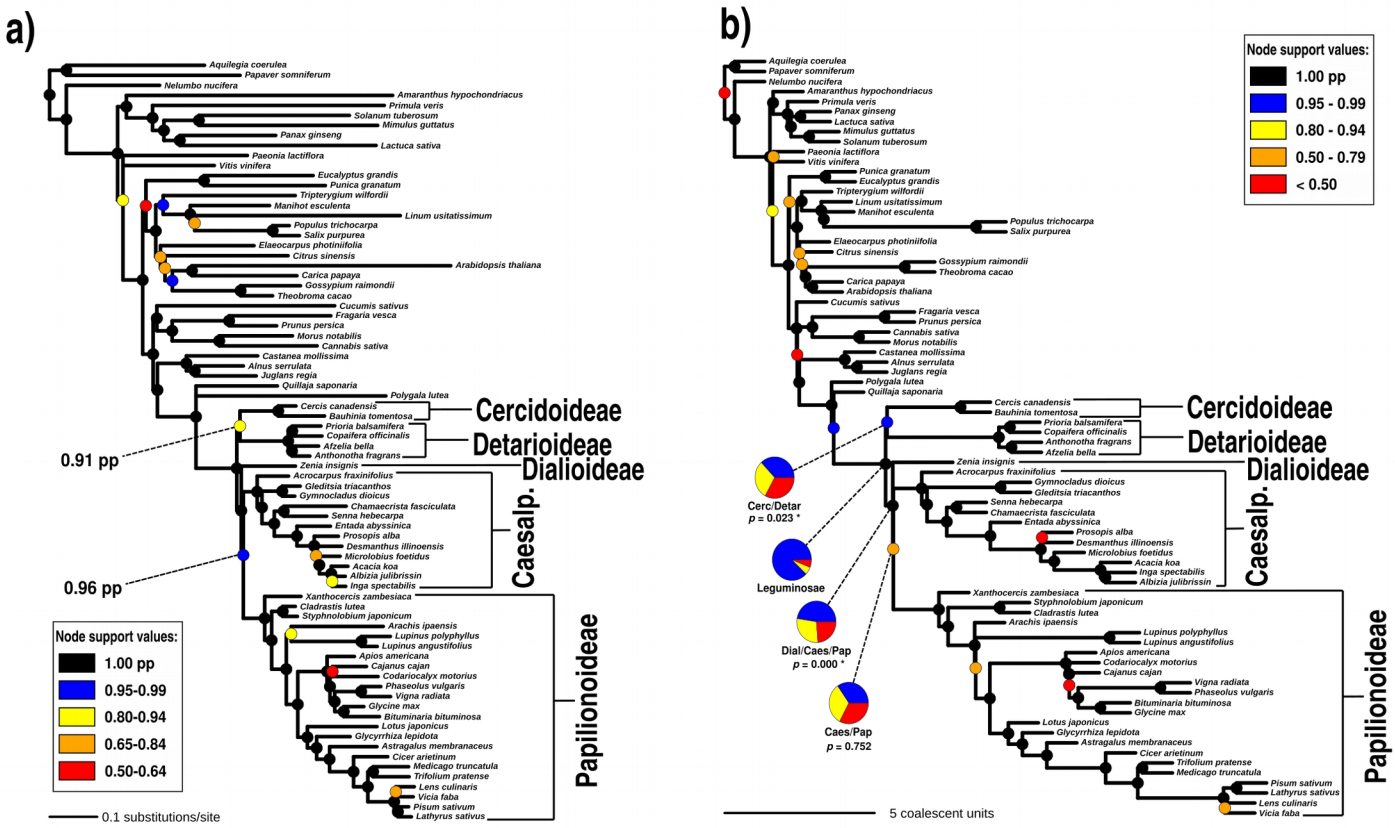


Figure 4. Bayesian and multi-species coalescent analyses yield congruent relationships, identical to those in Fig. 3 obtained with ML analyses on nuclear data. (a) Bayesian gene jackknifing majority-rule consensus tree of concatenated alignments of c. 220 genes per replicate, support indicated with coloured symbols on nodes represents posterior probability averaged over 25 replicates for 500 posterior trees each (in total 12,500 posterior trees). (b) Phylogeny estimated under the multi-species coalescent with ASTRAL from ML gene trees, support indicated with coloured symbols on nodes represents local posterior probability. Pie charts show relative quartet support for the first (blue) and the two (yellow and red) alternative quartets. *P*-values for the polytomy test are given for nodes B, E and F (see Fig. 3) below the respective pie charts for those nodes, significance (*p*-value ≤ 0.05) is indicated with an asterisk. See Figures S6 and S7 for phylogenetic trees with all posterior probability and quartet support values indicated.

The SplitsTree network (Fig. 5) shows relationships that are largely in line with the nuclear species tree, but is not entirely tree-like, including along the backbone of the family where edge lengths are shorter than elsewhere in the network.

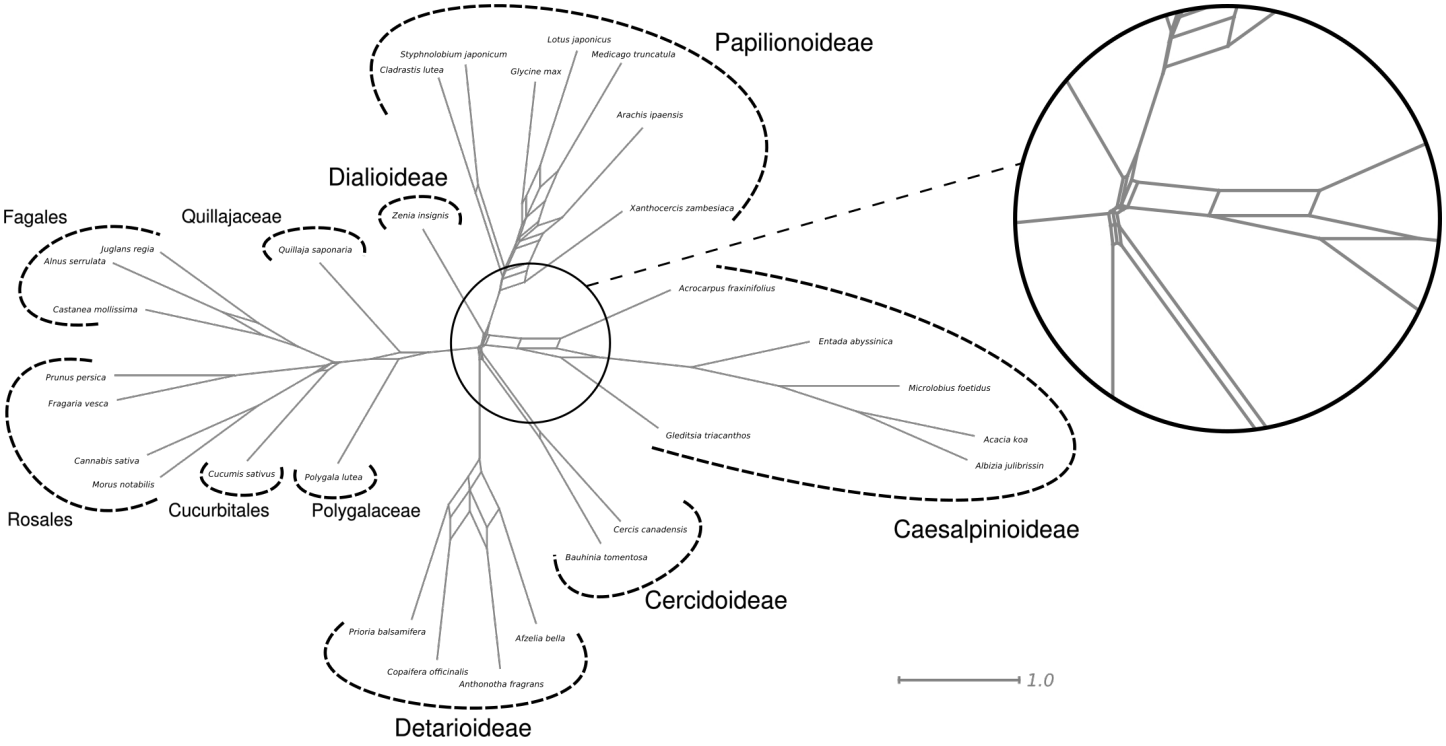


Figure 5. Filtered supernetwork inferred from the 1,103 1-to-1 orthologs, with extremely short internal edges around the origin of the legumes, highlighting their near-simultaneous divergence.

Evaluation of gene tree support and conflict

While the chloroplast and nuclear phylogenies show a different topology with regard to the first two dichotomies within legumes, all the ML, Bayesian and multi-species coalescent analyses of the nuclear data yield the same topology at the base of the family (Figs 3 & 4), showing a sister-group relationship of Cercidoideae and Detarioideae and a clade comprising the remaining three sampled subfamilies as their sister clade. Because the nuclear data set comprises 1,103 unlinked loci sampled from across the nuclear genome, while the recombination-free chloroplast genome constitutes just a single locus, the nuclear topology

should be considered as a more realistic estimate of species tree topology. However, when evaluating gene tree conflict, it is clear that many conflicting bipartitions exist, with the most prevalent being nearly as frequent across gene trees as compatible bipartitions (pie charts in Fig. 3). The quartet support calculated by ASTRAL is also low (37%, with alternative quartet supports 33% and 30%; pie charts in Fig. 4b). The relationships among the remaining three sampled subfamilies are also supported by significantly fewer bipartitions and lower quartet support than for example the legume crown node (pie charts in Figs 3 & 4b).

Rather than relying solely on ICA and quartet support values, we sought to evaluate in a more intuitive way how much support and conflict there is among gene trees for the deepest divergences in the legume family. For nodes labelled A-H in Figure 3, we counted how often a bipartition that is equivalent to that node in the species tree is encountered across gene trees, and how often those bipartitions received at least 50 or 80% BS. We did this on all RT homologs (n=7,621) in which all subfamilies and the outgroup were represented by at least one taxon, leading to 3,473 gene trees being considered. This shows that the legume family as a whole (node A), and the four subfamilies for which more than one taxon was sampled (nodes C, D, G and H), are all found to be monophyletic across the majority of gene trees (Fig. 6a & Table S3), and those bipartitions mostly receive at least 50 or 80% BS (Fig. 6a). Nodes B, E and F, that is, the relationships among the subfamilies, are recovered in many fewer gene trees, especially when considering only bipartitions with at least 50 or 80% BS. Expressing these differences in percentages of the total number of gene trees (n=3,473), this difference becomes especially stark, with nodes A, C, D, G and H receiving at least 80% BS in 33.14% – 74.43% of gene trees, while the same level of BS is found in only 1.38%, 2.62% and 1.21% of gene trees for nodes B, E and F, respectively. For these latter three nodes, we checked how often the most important conflicting bipartitions were present (Figs 6b-d & Table S3). These conflicting bipartitions are each less prevalent than those found by the concatenated ML and Bayesian analyses as well as by ASTRAL. This confirms that the recovered topology represents the relationships among legume subfamilies that is supported by the largest fraction of the genomic data used here, despite lack of phylogenetic signal across these nodes (Fig. 3) and significant and well-supported gene tree conflict, especially surrounding the root of the legumes (Fig. 6b).

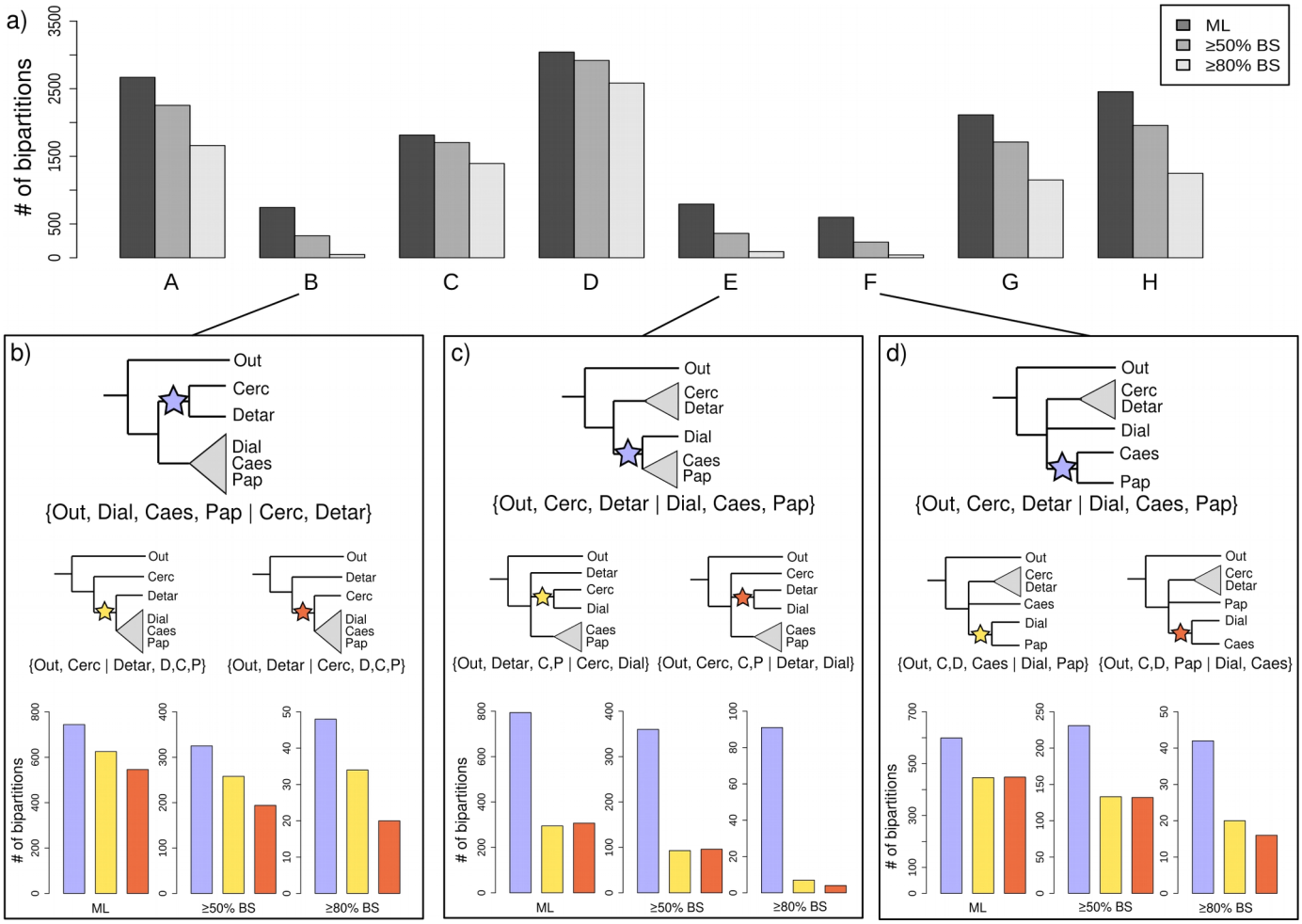


Figure 6. Leguminosae and its subfamilies are each supported by a large fraction of gene trees, in contrast to relationships among the subfamilies. (a) Prevalence of bipartitions that are equivalent to nodes A-H (see Fig. 3), among the 3,473 gene trees inferred from the RT homolog clusters (including 1-to-1 orthologs) in which all five subfamilies and the outgroup were included. Numbers of bipartitions are shown as counted from the best-scoring ML gene trees as well as taking only bipartitions with ≥50% and ≥80% bootstrap support (BS) into account, as indicated in the legend. (b-d) Prevalence of bipartitions for nodes B, E and F plotted next to the most common alternative bipartitions. The locations of the stars in the illustrations indicate the internodes of the phylogeny that are equivalent to the bipartitions for which counts are plotted below, as counted from the ML estimates and for bipartitions with ≥50% or ≥80% BS. Colours of the stars correspond to the colours of the bars in the barplots.

Discussion

Resolving the deep-branching relationships in the Leguminosae

Previous phylogenetic studies aimed at resolving deep relationships in legumes have relied on only a few chloroplast markers (Doyle et al., 1997; Wojciechowski et al., 2004; Lavin et al., 2005; Bruneau et al., 2008; LPWG, 2017), but here we show that even 72 protein-coding genes from the chloroplast genome fail to consistently resolve the root node with high support (Fig. 2c). Furthermore, substitution rate variation as evident from branch length disparity among legume subfamilies (Fig. 2b) (as previously shown for *matK* and *rbcL* by Lavin et al. (2005)), implies that while the chloroplast genome may be a useful marker to resolve relationships within Papilionoideae (particularly within the 50Kb-inversion clade), it is of limited use in other subfamilies, particularly Caesalpinioideae (Fig. 2a). Clearly, moving beyond the chloroplast genome and analysing nuclear gene data is necessary to improve phylogenetic hypotheses for the legume family, as found for other parts of the plant tree of life where chloroplast data have proven insufficiently informative (e.g. to resolve Mesangiosperms; Moore et al., 2010; Li et al., 2019). Nuclear data are also essential to detect ILS and/or introgression. Using nuclear gene data, we recovered a best-supported topology for the subfamily relationships that is different from the weakly supported chloroplast topology, and also quantify the strength of phylogenetic signal for alternative topologies.

Our results show that the difficulty of obtaining resolution for deep divergences in the legume family is in part caused by lack of phylogenetic signal in the chloroplast genome and a large fraction of the sampled nuclear genes (pie charts in Fig. 3), with too few substitutions having accumulated along the deepest short internodes, leading to only a small fraction of the gene trees showing strong support (≥80% BS) for relationships among these (Fig. 6 & Table S3). However, for a significant proportion of those genes that do have sufficient phylogenetic signal, we find strongly supported conflicting evolutionary histories. Putting aside methodological issues such as poor orthology inference for a number of genes, this conflict is likely to be caused by incomplete lineage sorting (ILS) (Pamilo & Nei, 1988; Maddison, 1997). Indeed, strong gene tree conflict caused by ILS is thought to be relatively common when

internodes are short due to rapid diversification and this provides an explanation as to why many relationships are contentious at all taxonomic levels (e.g. Pollard et al., 2006; Suh et al., 2015; Moore et al., 2017; but see Scornavacca & Galtier (2017) and Richards et al. (2018)).

Taken together, this could suggest that a fully bifurcating tree is not a good representation of the initial radiation of the legumes. As we show here, genes have many different evolutionary histories across the early divergences of legumes (Table S3), while the species tree merely represents the dominant evolutionary history. In the case of complete lack of phylogenetic signal, or equally prevalent conflicting evolutionary histories without a single dominant one, this would constitute a hard polytomy, implying (nearly) instantaneous divergence of three or more lineages, as demonstrated for Neoaves (Suh, 2016). In the legumes, there does appear to be one dominant evolutionary history in the relationships among subfamilies supported by a larger fraction of gene trees (Fig. 6), suggesting that the deep-branching relationships can be represented by a fully bifurcating topology. A hard polytomy at the root node of the legumes is also rejected by ASTRAL, but the same test did not reject a polytomy among Dialioideae, Caesalpinioideae and Papilionoideae. This is surprising since the relationships among these have been recovered in previous studies (Bruneau et al., 2008; LPWG, 2017), and are recovered here in all our analyses (Figs 2-4 & S1-S7) with high support in most of these (Figs 2c, 4a & S1-S4, S6). The bipartition counts (Fig. 6d) also suggest that a hard polytomy can likely be rejected for the relationships among these three subfamilies. However, the ICA value for a sister-group relationship of Caesalpinioideae and Papilionoideae is lower than for Cercidoideae and Detarioideae (0.70 vs 0.85) and support is even weaker in the ASTRAL analysis (0.58 pp). Since the levels of conflict are similar to that for the root of the legumes (Figs 6b & d), the lower support and failure to reject a polytomy may be caused by deeper gene coalescences than for the Cercidoideae/Detarioideae clade and/or introgression via hybridization shortly after divergence. Perhaps with denser taxon sampling, in particular for Dialioideae for which only one species was sampled here, it will be possible to reject a hard polytomy across this clade.

A further complication potentially affecting phylogeny reconstruction is the occurrence of whole genome duplications (WGDs) in the early evolution of the legumes (Cannon et al., 2015; Stai et al., 2019). This could lead to issues with ortholog detection or artefactual

inferences due to sub- or neofunctionalization of paralog copies independently in different lineages in the case of shared polyploidy among (some of) the subfamilies. However, Cannon et al. (2015) and Stai et al. (2019) inferred that only independent WGDs occurred in each of the subfamilies, suggesting that these issues should have minimal effect on the relationships among subfamilies. While the homolog trees inferred here during the ortholog selection procedure are suitable to test the placements of WGDs on the phylogeny, this is beyond the scope of this study and will be addressed elsewhere (Koenen et al., in prep.).

Implications for our understanding of the evolution of legume diversity and traits

Regardless of whether hard polytomies can be rejected or not, the lack of phylogenetic signal and significant conflict among gene trees at the base of the legumes are indicative of rapid successive divergences. This near-simultaneous divergence of the six main lineages of legumes is highly relevant to our understanding of the evolution of legume diversity and the appearance of key traits. Over the last few decades, the prevailing characterization of legume evolution has been that of mimosoids and papilionoids as derived clades that evolved from a paraphyletic grade of caesalpinioid legumes (e.g. LPWG, 2013a). This led to the misplaced characterization of several caesalpinioid lineages as in some way “basal” or “early-diverging” (see LPWG, 2013a and references therein). Such characterisations are commonly made, but are in fact phylogenetic misinterpretation, given that basal nodes are ancestral nodes and at each bifurcating node two sister-groups diverge from each other concurrently, neither of them earlier (Crisp & Cook, 2005). Species-poor successive sister-groups of species-rich clades are often mistakenly referred to as basal or early-diverging, and this appears also to have been the case in legumes, where the mimosoids and papilionoids have (vastly) more species than other lineages such as Cercidoideae, Detarioideae, Duparquetioideae and Dialioideae. This can lead to the erroneous assumption that lineages such as Cercidoideae, Detarioideae, Duparquetioideae and Dialioideae have retained more ancestral traits than the species-rich mimosoid and papilionoid clades (Crisp & Cook, 2005).

Moreover, we show that the branching order among the subfamilies is rather insignificant, with short internodes, conflicting relationships across gene trees and long stem

lineages subtending each subfamily, which is particularly well visualized in the SplitsTree supernetwork (Fig. 5). This evidence for near-simultaneous divergence of the six subfamilies provides an additional argument to abandon the idea of “early-diverging” lineages in legumes. Given that the stem lineage of the family is also rather long, most trait evolution likely occurred along the long legume and subfamily stem branches, rather than derived legume traits having evolved in a stepwise fashion across the first divergences in the family. In comparative analyses, the branching order among subfamilies is unlikely to be meaningful and it should be effectively considered as a polytomy with respect to trait evolution. We therefore suggest that typical legume traits evolved along the stem lineage of the family and were shared by the earliest stem-relatives of each subfamily, i.e. the earliest stem-relatives of each subfamily probably had similar traits.

Over the past decade, there was a strong debate regarding into how many and which subfamilies the legumes should be classified (LPWG 2013b, 2017). That legumes seem to consist of six (nearly) simultaneously originating lineages strongly supports the outcome of the debate which resulted in the recognition of six legume subfamilies (LPWG, 2017).

This view of near simultaneous divergence of subfamilies also suggests that many of the traits shared across legume subfamilies (LPWG (2017): Table 1) could be plesiomorphic, having been independently lost or modified in some subfamilies and retained in others. An alternative hypothesis is that these traits are not ancestral to all legumes and have evolved independently in different lineages, leading to homoplasy. Somewhat intermediate is the hypothesis of a shared cryptic precursor trait that can lead to deep homology, where similar traits evolved independently from a shared genetic basis (Shubin et al., 2009; Scotland, 2010). For instance, this could potentially explain the homoplasious distribution of extra-floral nectaries across legumes (Marazzi et al., 2012), which are present in several subfamilies but are different in structure and location, casting doubt on a single origin and prompting the possibility of a shared genetic precursor (Marazzi et al., 2012 & in press).

However, the precursor trait hypothesis may be motivated more by the notion that massive parallel loss of a trait is less parsimonious than assuming a few more independent gains. For instance, the evolution of nitrogen fixation in root nodules, a trait that is especially prominent in legumes, has been suggested to be driven by a cryptic precursor trait in the

nitrogen-fixing clade of angiosperms (Werner et al., 2014), with five independent gains in legumes, within subfamilies Caesalpinioideae and Papilionoideae, being most parsimonious (Doyle, 2016). Recent genomic evidence, however, supports the scenario of a single origin shared by the whole of the nitrogen-fixing clade of angiosperms, with massive parallel losses in each of the four subclades (the orders Cucurbitales, Fabales, Fagales and Rosales) of the nitrogen-fixing clade (van Velzen et al., 2018a; Griessmann et al., 2018). Such a scenario suggests that the legume ancestor was also a nodulator, and given the rapid successive speciation associated with the initial divergence of legumes documented here, that stem relatives of all subfamilies likely also had the ability to nodulate, but that nodulation was presumably lost in parallel along the long stem lineages or early in the crown group divergences of Cercidoideae, Detarioideae, *Duparquetia* and Dialioideae, in which no nodulating species are known. Finding out when and why nodulation has apparently been lost in all but two of the legume subfamilies will be important for understanding the causes of massive parallel loss of nodulation in the nitrogen-fixing clade of angiosperms (van Velzen, 2018b).

Examples of other traits that are either plesiomorphic or homoplasious among and/or within subfamilies include: wood with vestured pits (also present in some Polygalaceae; Jansen et al., 2001) and absent in Cercidoideae, *Duparquetia* and most Dialioideae (LPWG, 2017)); ectomycorrhizal symbiosis (known to occur in Detarioideae, Caesalpinioideae and Papilionoideae (Smith et al., 2011)); and floral symmetry which is variable across all non-monotypic subfamilies (Cardoso et al., 2013; LPWG, 2017; Ojeda et al., 2019). These and other traits are candidates for comparative (genomic) analyses based on the new phylogenetic framework presented here, to test the hypothesis that several legume traits are ancestral with multiple independent losses rather than independent gains.

Finally, it is clear that the lack of resolution among the six legume subfamilies is also relevant for inferring the placements of WGDs and reconstructing the ancestral legume genome. For example, the recent suggestion by Stai et al (2019) that *Cercis* could represent the genome duplication status of the ancestral legume, is in part based on placement of Cercidoideae as sister to the rest of the legumes, which we show is poorly supported in the chloroplast alignment and not the most likely species tree topology based on nuclear genes.

Concluding remarks

In this study, we present some of the first phylogenetic analyses using genome-scale data for the Leguminosae, sampling representatives of all six subfamilies. While our results show overwhelming support for monophyly of the family and each of the five non-monotypic subfamilies, there is both a paucity of phylogenetic signal and strongly conflicting signals across gene trees regarding relationships among them. This suggests that the six main lineages of legumes originated in quick succession, or nearly simultaneously, with significant implications for understanding the evolution of legume diversity and traits.

We also show that it is essential in phylogenomic studies to explicitly evaluate conflicting phylogenetic signals across the genome. By taking into account alternative topologies with high BS across gene trees (Fig. 6), the phylogenomic complexity of the initial radiation of the legumes is revealed. More generally, this study adds to an increasing understanding of the limits to phylogenetic resolution, highlighting the role of rapid successive deep divergences in causing lack of phylogenetic signal and gene tree conflict across the Tree of Life.

Acknowledgements

This work was supported by the Swiss National Science Foundation (Grant 31003A_135522 to CEH), the Department of Systematic and Evolutionary Botany, University of Zurich, the Natural Sciences and Engineering Research Council of Canada (Grant to AB), the U.K. National Environment Research Council (Grant NE/1027797/1 to RTP) and the Fonds de la Recherche Scientifique of Belgium (Grant J.0292.17 to OH). We thank the S3IT of the University of Zurich for use of the ScienceCloud computational infrastructure and the Functional Genomics Center Zurich (FGCZ) for library preparation and sequencing.

Author contribution

EK and CEH designed the research; EK carried out the research and wrote the manuscript; all authors contributed to data analysis, collection and/or interpretation, as well as to writing of the final version of the manuscript.

References

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin V.M, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19: 455–477.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bruneau A, Mercure M, Lewis GP, Herendeen PS. 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. *Botany* 86: 697–718.

Cardoso D, de Queiroz LP, Pennington RT, de Lima HC, Fonty E, Wojciechowski MF, Lavin M. 2012. Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. *American Journal of Botany* 99: 1991–2013.

Cardoso D, Pennington RT, de Queiroz LP, Boatwright JS, Van Wyk B-E, Wojciechowski MF, Lavin M. 2013. Reconstructing the deep-branching relationships of the papilionoid legumes. *South African Journal of Botany* 89: 58–75.

Crepet WL, Herendeen PS. 1992. Papilionoid flowers from the early Eocene of southeastern North America. In: Herendeen PS, Dilcher DL, eds. *Advances in legume systematics part 4: The fossil record*. Richmond, UK: Royal Botanic Gardens, Kew, 43–55.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology* 10: 210.

Crisp MD, Cook LG. 2005. Do early branching lineages signify ancestral traits? *Trends in Ecology and Evolution* 20: 122-128.

Doyle JJ. 1995. DNA data and legume phylogeny: a progress report. In: Crisp MD and Doyle JJ, eds. *Advances in Legume Systematics part 7: Phylogeny.* Richmond, UK: Royal Botanic Gardens, Kew, 11–30.

Doyle JJ. 2016. Chasing unicorns: Nodulation origins and the paradox of novelty. *American Journal of Botany* 103: 1865-1868.

Doyle JJ, Doyle JL, Ballenger JA, Dickson EE, Kajita T, Ohashi H. 1997. A phylogeny of the chloroplast gene *rbc*L in the Leguminosae: taxonomic correlations and insights into the evolution of nodulation. *American Journal of Botany* 84: 541-554.

Dugas DV, Hernandez D, Koenen EJ, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT and Hajrah NH. 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. *Scientific reports* 5*:* 16958.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* 29: 644–652.

Griesmann M, Chang Y, Liu X, Song Y, Haberer G, Crook M., Billault-Penneteau B, Lauressergues D, Keller J, Imanishi L, Roswanjaya YP. 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* 361: p.eaat1743.

Jansen S, Baas P, Smets E. 2001. Vestured pits: their occurrence and systematic importance in eudicots. *Taxon* 50: 135-167.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34: 772–773.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* 21: 1095–1109.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology* 62: 611–615.

Lassmann T, Hayashizaki Y, Daub CO. 2009. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25: 2839–2840.

Lavin M, Herendeen PS, Wojciechowski MF. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Systematic Biology* 54: 575–594.

Le Q, Dang C, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution* 29: 2921–2936.

Lewis GP, Schrire B, Mackinder B, Lock M. 2005. *Legumes of the World*. Richmond, UK: Royal Botanic Gardens Kew.

Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, Gitzendanner MA, Fritsch PW, Cai J, Luo Y. et al. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* 5: 461-470.

LPWG (Legume Phylogeny Working Group). 2013a. Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species-rich clades. *Taxon* 62: 217–248.

LPWG (Legume Phylogeny Working Group). 2013b. Towards a new classification system for legumes: Progress report from the 6th International Legume Conference. South African Journal of Botany 89: 3–9.

LPWG (Legume Phylogeny Working Group). 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* 66: 44–77.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, Gray JC. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* 20: 1700-1710.

Manzanilla V, Bruneau A. 2012. Phylogeny reconstruction in the Caesalpinieae grade (Leguminosae) based on duplicated copies of the sucrose synthase gene and plastid markers. *Molecular Phylogenetics and Evolution* 65: 149-162.

Marazzi B, Ané C, Simon MF, Delgado-Salinas A, Luckow M, Sanderson MJ. 2012. Locating evolutionary precursors on a phylogenetic tree. *Evolution* 66: 3918-3930.

Marazzi B, Gonzalez AM, Delgado-Salinas A, Luckow MA, Ringelberg J, Hughes C.E. 2019. Extrafloral nectaries in Leguminosae: phylogenetic distribution, morphological diversity and evolution. *Australian Systematic Botany*, in press.

McKey D. 1994. Legumes and nitrogen: the evolutionary ecology of a nitrogen-demanding lifestyle. In: Sprent JI, McKey D. eds. *Advances in Legume Systematics 5: The Nitrogen Factor.* Richmond, U.K.: Royal Botanic Gardens, Kew, 211-228.

Mirarab S, Reaz R, Bayzid MdS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: i541–i548.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Acadamy of Sciences, USA* 107: 4623–4628.

Moore AJ, Vos JMD, Hancock LP, Goolsby E, Edwards EJ. 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the portullugo clade (Caryophyllales). *Systematic biology* 67: 367-383.

Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ. 2013. Heterogeneous models place the root of the placental mammal phylogeny. *Molecular Biology and Evolution* 30: 2145-2156.

Ojeda D.I, Koenen E, Cervantes S, de la Estrella M, Banguera-Hinestroza E, Janssens SB, Migliore J, Demenou B, Bruneau A, Forest F, Hardy OJ. 2019. Phylogenomic analyses reveal an exceptionally high number of evolutionary shifts in a florally diverse clade of African legumes. *Molecular Phylogenetics and Evolution* 137: 156-167.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568–583.

Pennington RT, Lavin M, Ireland H, Klitgaard B, Preston J, Hu JM. 2001. Phylogenetic relationships of basal papilionoid legumes based upon sequences of the chloroplast trnL intron. *Systematic Botany* 26: 537-557.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS Genetics* 2: p.e173.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS ONE 6: e22594.

Richards EJ, Brown JM, Barley AJ, Chong RA, Thomson RC. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological?. *Systematic Biology* 67: 847-860.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798.

Rokas A, Carroll SB. 2006. Bushes in the Tree of Life. *PLoS Biology* 4: e352.

Romiguier, J, Ranwez, V, Delsuc, F, Galtier, N, Douzery, E.J. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular Biology and Evolution* 30: 2134-2144.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327–331.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668.

Sayyari E, Mirarab S. 2018. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* 9: 132.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863–864.

Scornavacca C, Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology* 66: 112-120.

Scotland RW. 2010. Deep homology: a view from systematics. Bioessays 32: 438-449.

Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* 1: 0126.

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. *Nature* 457: 818.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Queinnec E, Ereskovsky A, Lapebie P. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology* 27: 958-967.

Simon MF, Grether R, de Queiroz LP, Skema C, Pennington RT, Hughes CE. 2009. Recent assembly of the Cerrado, a Neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proceedings of the National Acadamy of Sciences, USA* 106: 20359–20364.

Smith ME, Henkel TW, Aime MC, Fremier AK, Vilgalys R. 2011. Ectomycorrhizal fungal diversity and community structure on three co-occurring leguminous canopy tree species in a Neotropical rainforest. *New Phytologist* 192: 699-712.

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

Stai JS, Yadav A, Sinou C, Bruneau A, Doyle JJ, Fernández-Baca D, Cannon SB. 2019. Cercis: a Non-Polyploid Genomic Relic within the Generally Polyploid Legume Family. *Frontiers in Plant Science* 10: 345.

Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology* 13: e1002224.

Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zoologica Scripta* 45: 50–62.

Sukumaran J, Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26: 1569–1571.

van Velzen R, Holmer R, Bu F, Rutten L, van Zeijl A, Liu W, Santuari L, Cao Q, Sharma T, Shen D, Roswanjaya Y. 2018a. Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proceedings of the National Academy of Sciences, USA* 115: E4700-E4709.

van Velzen R, Doyle JJ, Geurts R. 2018b. A Resurrected Scenario: Single Gain and Massive Loss of Nitrogen-Fixing Nodulation. Trends in Plant Science 24: 49-57.

Werner GD, Cornwell WK, Sprent JI, Kattge J, Kiers ET. 2014. A single evolutionary innovation drives the deep evolution of symbiotic N 2-fixation in angiosperms. *Nature Communications* 5: 4087.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* 111: E4859-E4868.

Williams AV, Boykin LM, Howell KA, Nevill PG, Small I. 2015. The complete sequence of the Acacia ligulata chloroplast genome reveals a highly divergent clpP1 gene. *PLoS One* 10: p.e0125768.

Wojciechowski MF, Lavin M, Sanderson MJ. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *American Journal of Botany* 91: 1846–1862.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel nrganisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 31: 3081–3092.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30: 614–620.

Supplementary Information (see Appendix III on page 217)

Table S1. Accession information for taxa included in the chloroplast alignment.

Table S2. Accession information for taxa included in the nuclear genomic and transcriptomic data set.

Table S3. Counts of bipartitions representing nodes A-H (Fig. 3) and conflicting bipartitions representing other subfamily relationships among 3,473 gene trees.

Figure S1. ML topology as inferred by RAxML from amino acid alignment of chloroplast genes under the LG4X model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

Figure S2. Bayesian majority-rule consensus tree inferred with Phylobayes from amino acid alignment of chloroplast genes under the CATGTR model. Numbers on nodes indicate posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

Figure S3. ML topology as inferred by RAxML from nucleotide alignment of chloroplast genes under the GTR + G model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

Figure S4. Bayesian majority-rule consensus tree inferred with Phylobayes from nucleotide alignment of chloroplast genes under the CATGTR model. Numbers on nodes indicate the posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

Figure S5. ML topology as inferred by RAxML from a concatenated alignment of 1,103 nuclear genes, under the LG4X model. Numbers on nodes indicate Internode Certainty All (ICA) values, as estimated from gene trees of the same 1,103 genes.

Figure S6. Bayesian gene jackknifing majority-rule consensus tree inferred with Phylobayes from a concatenated alignment of 1,103 nuclear genes. Numbers on nodes indicate posterior probabilities (pp), averaged over 500 posterior trees each, for 25 replicates (12,500 posterior trees in total).

Figure S7. Phylogeny estimated under the multi-species coalescent summary method with ASTRAL. Support values indicated represent local posterior probability (blue rectangles) and quartet support (yellow rectangles).

Chapter II

The Origin and Early Evolution of the Legumes are a Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-Paleogene (K-Pg) Boundary

Authors:
Erik J.M. Koenen[1], Dario I. Ojeda[2,3], Freek T. Bakker[6], Jan J. Wieringa[7], Catherine Kidner[8,9], Olivier Hardy[2], R. Toby Pennington[8,10], Patrick S. Herendeen[11], Anne Bruneau[4] and Colin E. Hughes[1]

[1] Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstrasse 107, CH-8008, Zurich, Switzerland
[2] Service Évolution Biologique et Écologie, Faculté des Sciences, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050, Brussels, Belgium
[3] Norwegian Institute of Bioeconomy Research, Høgskoleveien 8, 1433 Ås, Norway
[4] Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
[5] Naturalis Biodiversity Center, Leiden, Darwinweg 2, 2333 CR, Leiden, The Netherlands
[6] Royal Botanic Gardens Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, U.K.
[7]School of Biological Sciences, University of Edinburgh, King's Buildings, Mayfield Rd, Edinburgh, UK
[8] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX4 4RJ, U.K.
[9] Chicago Botanic Garden, 1000 Lake Cook Rd, Glencoe, IL 60022, U.S.A.
[10] Institut de Recherche en Biologie Végétale and Département de Sciences Biologiques, Université de Montréal, 4101 Sherbrooke St E, Montreal, QC H1X 2B2, Canada

CHAPTER II

Abstract – The consequences of the Cretaceous-Paleogene (K-Pg) boundary (KPB) mass extinction for the evolution of plant diversity remain poorly understood, even although evolutionary turnover of plant lineages at the KPB is central to understanding the assembly of the Cenozoic biota. One aspect that has received considerable attention is the apparent concentration of whole genome duplication (WGD) events around the KPB, which may have played a role in survival and subsequent diversification of plant lineages. In order to gain new insights into the origins of Cenozoic biodiversity, we examine the origin and early evolution of the legume family (Leguminosae or Fabaceae), which with c. 20.000 species, is the third largest family of Angiospermae, that rose to prominence after the KPB and for which multiple WGD events are hypothesized to have occurred early in its evolution. Using a recently inferred phylogenomic framework for the family, we investigate the placements of WGDs during the early evolution of the legumes using gene tree reconciliation methods, gene count data and phylogenetic supernetwork reconstruction. Using a set of 20 fossil calibrations we estimate a revised timeline of legume evolution based on 36 genes selected as informative and evolving in an approximately clock-like fashion. To establish the timing of WGDs we also date duplication nodes in gene trees. Our results suggest that either a pan-legume WGD event occurred on the stem lineage of the family, or an allopolyploid event occurred along the backbone of the family, with additional nested WGDs subtending subfamilies Papilionoideae and Detarioideae. Gene tree reconciliation methods that do not account for allopolyploidy may be misleading in inferring an earlier WGD event at the time of divergence of the two parental lineages of the polyploid, suggesting that the allopolyploid scenario is more likely. We show that the crown age of the legumes dates back to the Maastrichtian or early Paleocene and that, apart from the detarioid WGD, paleopolyploidy occurred close to the KPB. We conclude that the origin and early evolution of the legumes followed a complex history, in which multiple auto- and/or allopolyploidy events coupled with rapid diversification are

associated with the mass extinction event at the KPB, ultimately underpinning the evolutionary success of the Leguminosae in the Cenozoic.

Keywords: Cretaceous-Paleogene (K-Pg) boundary, Leguminosae, Fabaceae, Whole Genome Duplication events, paleopolyploidy, allopolyploidy, phylogenomics

The Cretaceous-Paleogene (K-Pg) boundary (KPB), 66 Million years ago (Ma), is defined by the mass extinction event that famously killed the non-avian dinosaurs and led to major turnover in the earth's biota. The Chicxulub meteorite impact is generally thought to have been the cause of the mass extinction, but Deccan trap flood basalt volcanism likely contributed or may have been the primary cause, in line with previous global mass extinctions that are all related to volcanism (Keller, 2014). The KPB event determined in significant part the composition of the Earth’s modern biota, because many lineages that were successful in repopulating the planet and diversifying in the wake of the KPB, have remained abundant and diverse throughout the Cenozoic until the present. Probably the best-known examples of successful post-KPB lineages are the mammals and birds, both inconspicuous elements of the Cretaceous fauna, while their core clades Placentalia and Neoaves became ubiquitous throughout Cenozoic fossil faunas. Plants were also severely affected by the KPB, with a clear shift in floristic composition and a drop in macrofossil species richness of up to 78% reported across boundary-spanning fossil sites in North-America (Wilf and Johnson, 2004; McElwain and Punyasena, 2007; Vajda and Bercovici, 2014). In addition, a global fungal spike followed by a global fern spike in the palynological record (Vajda et al., 2001; Barreda et al., 2012) are consistent with sudden KPB ecosystem collapse and a recovery period characterized by low diversity vegetation dominated by ferns. Although the KPB is not considered a major extinction event for plants, as no plant family appears to have been lost at the KPB (McElwain and Punyasena, 2007; Cascales-Miñana and Cleal, 2014), a sudden increase in net diversification rate in the Paleocene has been inferred from a

large paleobotanical data set (Silvestro et al., 2015), suggesting increased origination following the KPB.

Macro-evolutionary dynamics of plant clades across the KPB extinction event have received less attention than prominent vertebrate clades, even though plants are the main primary producers and structural components of terrestrial ecosystems, such that the shaping and diversification of the Cenozoic biota cannot be fully understood without understanding the consequences of the KPB for evolutionary turnover of plant diversity. One aspect of plant evolution in relation to the KPB that has been investigated is the apparent concentration of whole genome duplication (WGD) events around the KPB (Fawcett et al., 2009; Vanneste et al., 2014; Lohaus and Van de Peer, 2016; but see Cai et al., 2019). This is explained by the idea that polyploid lineages could have had enhanced survival and establishment across the KPB (Lohaus and Van de Peer, 2016) as well as greater potential to diversify rapidly thereafter relative to diploids (Levin and Soltis, 2018). Recent work is revealing the prevalence and significance of WGDs in shaping the evolution of the flowering plants (Wendel, 2015; Soltis et al., 2016; Yang et al., 2018; Cai et al., 2019; Conover et al., 2019). Determining the phylogenetic placements of WGDs and estimating their timing is a central issue in plant systematics, but remains challenging, with different lines of evidence yielding conflicting results, such that many WGDs remain contentious and poorly understood (e.g. Conover et al., 2019).

We examine the role of the KPB in shaping Cenozoic plant diversity by investigating the origin and early evolution of the legume family, including the placement and timing of early legume WGDs. The legume family (Leguminosae or Fabaceae), perhaps more than any other plant clade, appears to parallel Placentalia and Neoaves. No clearly identifiable legume fossils are known that pre-date the KPB (Herendeen and Dilcher, 1992), but the family was already abundant and diverse in one of the earliest examples of modern type rainforests in the Paleocene (Wing et al., 2009; Herrera et al., in press). The oldest known fossils that are clearly referable to (stem groups of) subfamilies are from close to the Paleocene-Eocene Thermal Maximum (PETM)

(morphotype # CJ76 of c. 58 Ma (Wing et al., 2009) can be referred to Caesalpinioideae and *Barnebyanthus buchananensis* of c. 56 Ma to Papilionoideae (Crepet and Herendeen, 1992)) and legumes are a ubiquitous element of many Eocene, Oligocene and Neogene floras (Herendeen and Dilcher, 1992). Today, it is the third most species-rich angiosperm family, and arguably the most spectacular evolutionary and ecological radiation of any angiosperm family (McKey, 1994). The six main lineages of the legume family, recently recognized as subfamilies in a new classification (LPWG, 2017), apparently diverged nearly instantaneously (Koenen et al., submitted), mirroring Placentalia (Teeling and Hedges, 2013) and Neoaves (Suh et al., 2015; Suh, 2016).

The apparent rapid diversification of the legumes soon after the KPB, and the occurrence of multiple WGDs during their early evolution (Cannon et al., 2015; Stai et al., 2019), make the family an excellent model to investigate the possible association of WGDs with the KPB. However, there is uncertainty about how many WGDs were involved in the early evolution of legumes and their phylogenetic placements. From whole genome sequencing studies, it is well established that several taxa in subfamily Papilionoideae share a WGD event (Mudge et al., 2005; Cannon et al., 2006). Using gene tree topologies, this Papilionoideae WGD has subsequently been shown to be most likely shared among the whole subfamily and not with any of the other subfamilies, for three of which independent WGDs where suggested (Cannon et al., 2015). More recently, independent WGDs were hypothesized to have occurred early in the evolution of the five subfamilies for which genomic data is available, where a WGD in Cercidoideae excludes the genus *Cercis*, the sister group to the rest of that subfamily (Stai et al., 2019). The more parsimonious explanations of a single WGD shared across all legumes, or across multiple subfamilies, remain to be tested using more fully sampled gene trees. Furthermore, understanding the placements of these early legume WGDs is complicated by the apparent near-simultaneous divergence of the six subfamilies (Koenen et al., submitted).

Uncertainty also surrounds the age of the legume family. While the legumes are not known with certainty from any Cretaceous fossil site, the family has a long stem lineage dating back to c. 80 – 100 Ma (Wang et al., 2009; Magallón et al., 2015). This long ghost lineage means that the timing of the initial radiation of the family, as well as of legume WGDs, and notably whether they pre- or post-date the KPB, are uncertain. In Placentalia and Neoaves, divergence time estimation has been much debated, with some studies using molecular sequence data for divergence time estimation suggesting that both clades originated and diversified well before the KPB, implying that many lineages of both clades survived the end-Cretaceous event (Cooper and Penny, 1997; Jetz et al., 2012; Meredith et al., 2011). However, like the legumes, both groups first appear in the Paleocene fossil record. A phylogenetic study of mammals combining both molecular sequence data and morphological characters to enable inclusion of fossil taxa, found only a single placental ancestor crossing the KPB (O'Leary et al., 2013; but see Springer et al., 2013; dos Reis et al., 2014). Alternatively, it has been argued that diversification of Placentalia followed a "soft explosive" model, with a few lineages crossing the KPB followed by rapid ordinal level radiation during the Paleocene (Phillips, 2015; Phillips and Fruciano, 2018). Recent time-calibrated phylogenies for birds showed the age of Neoaves to also be close to the KPB (Jarvis et al., 2014; Claramunt and Cracraft, 2015; Prum et al., 2015), with initial rapid post-KPB divergence represented by a hard polytomy (Suh, 2016). For legumes, it is similarly unlikely that the modern subfamilies have Cretaceous crown ages. These clades, in particular Papilionoideae, Caesalpinioideae and Detarioideae, appear to have rapidly diversified following their origins, which would imply mass survival of large numbers of legume lineages across the KPB. Furthermore, diversification into the six main lineages of legumes appears to have occurred rapidly (Lavin et al., 2005), indeed nearly simultaneously (Koenen et al., submitted), with long stem branches leading to each of the modern subfamilies. Therefore, two hypotheses seem plausible: (1) the legumes have a Cretaceous crown age and diversified into the six subfamilies prior to the KPB, while crown radiations of

the subfamilies occurred (shortly) after the mass extinction event, corresponding to a "soft explosive" model, or (2) a single legume ancestor crossed the KPB and rapidly diversified into six main lineages in the wake of the mass extinction event, corresponding to a "hard-explosive" model, with the subsequent subfamily radiations related to the Paleocene-Eocene Thermal Maximum (PETM) and/or Eocene climatic optimum. Currently available molecular crown age estimates for the family range from c. 59 to 64 Ma (Lavin et al., 2005; Bruneau et al., 2008; Simon et al., 2009). These studies, however, lacked extensive sampling of outgroup taxa and relied instead on fixing the stem age of the legumes, thereby compromising the ability to estimate the crown age. Furthermore, these earlier studies relied exclusively on chloroplast sequences, for which evolutionary rates are known to vary strongly across legumes (Lavin et al., 2005), such that nuclear gene data are likely to be better suited for estimating divergence times (Christin et al., 2014).

In this study, we evaluate the number of WGDs during the early evolution of the legumes and whether one or more of them are shared across multiple subfamilies. We use gene tree reconciliation methods to identify the most likely placements of WGDs along the legume backbone and test these placements with a probabilistic method using gene count data. We also evaluate the possibility of allopolyploidy involving one or more lineages with phylogenetic supernetwork reconstruction and gene tree reconciliation with multi-labeled (MUL) trees. Secondly, we evaluate whether the origin of the legumes and WGDs during the early evolution of the family are closely associated with the KPB by inferring a new legume chronogram based on 36 informative and relatively clock-like nuclear genes and 20 fossil calibration points, as well as assessing the timing of duplication nodes in gene trees. More generally, this study addresses important questions surrounding the links between and consequences of WGDs and the KPB for the evolution of Cenozoic flowering plant diversity and the complications of phylogenomic inferences in deep time when paleopolyploidy is coupled with rapid successive divergences.

CHAPTER II

MATERIAL & METHODS

Gene Tree Inference

We used sets of homolog clusters generated prior to extracting orthologs for species tree inference using the Yang and Smith (2014) pipeline, based on a nuclear gene dataset derived from genome and transcriptome sequences for representatives of five of the six legume subfamilies and a large set of eudicot outgroups assembled by Koenen et al. (submitted). Taxon occupancy for each of the analyses described below is included in Table S1. These homolog clusters include multiple sequences per taxon representing paralogs for non-terminal gene duplications, such that duplications that are restricted to a terminal taxon are not included. The clusters were aligned with MAFFT v. 7.187 (Katoh and Standley, 2013) using the G-INSi algorithm. All sites with more than 5% missing data were removed with BMGE (Criscuolo et al., 2010) and all sequences with more than 75% gaps were removed, to avoid having fragmented paralog sequences present, which could inflate the number of gene duplications. These data removal steps also led to the elimination of clusters with large amounts of missing data. Tree estimation was then repeated on these clusters, with RaxML v. 8.2 (Stamatakis, 2014) using the WAG + G model and 100 rapid bootstrap replicates.

Mapping of Gene Duplications

From the homolog trees, we extracted rooted clades to use as input gene trees for gene duplication mapping analysis with Phyparts (Smith et al., 2015). This method counts for each node the number of gene trees in which at least two descendent taxa are represented by at least two paralogous sequences. *Aquilegia* and *Papaver* were used as outgroup taxa to root and extract the paralog clades. Phyparts was run with and without a 50% bootstrap cut-off.

Apart from the relatively simple Phyparts method, we performed gene tree reconciliation with a model of gene duplication and loss (horizontal transfers were not considered) using Notung v 2.9 (Stolzer et al., 2012) on the rosid portion of the species tree. Notung can account for ILS when using non-binary trees (i.e. trees with polytomies), therefore we introduced three polytomies for unsupported nodes in the species tree (at the base of Fabales and two small clades within Caesalpinioideae and Papilionoideae). Additionally, an analysis was run with the complete legume backbone collapsed to a polytomy, since ILS has likely occurred among the first divergences in the family (Koenen et al., submitted). The input gene trees were extracted from the homolog clusters as for the Phyparts analysis, but with all non-rosid taxa as the outgroup, such that the older Pentapetalae hexaploidization is not included in the analysis. First, we used the --rearrange option in Notung with an 80% bootstrap threshold such that poorly supported branches in the gene trees are rearranged according to the relationships found in the species tree. This has the drawback that in the case of missing data or duplicate gene loss, some genuine gene duplications with lower support are reconciled to a more inclusive clade. However, without performing this rearrangement step, we found that many more gene duplications were inferred across all nodes in the tree, presumably in part caused by gene tree estimation errors. Next, we ran the reconciliation analysis in --phylogenomics mode and analysed the number of inferred duplications on each node of the tree. We set the cost of duplications at 10, and the cost of gene loss at 0.1 to avoid missing data from transcriptomes adding much to the reconciliation score. We also explored a range of other settings but the results did not change significantly.

Testing Placements of WGDs using Gene Count Data

We used the WGDgc package in R (Rabier et al., 2013) to test the hypothesized placements of WGDs from the Phyparts and Notung results. This is a probabilistic

method where the background gene duplication and loss rates are modelled by a birth and death process, while WGDs are added on specific branches of the species tree. Both the birth-death rates and the duplicate gene retention rates for WGDs are estimated with maximum likelihood and the likelihood of different configurations of WGDs on the species tree can be compared. We extracted gene count data from the rosid gene trees that were used in the Notung analysis, but removed *Eucalyptus grandis* and *Punica granatum* in order to have two large clades at the root. Several transcriptome accessions with relatively high levels of missing data were removed. The count data were filtered to include at least one copy in both main clades at the root and at least one copy in each of the five sampled legume subfamilies. Analyses were run on several different models with two, three or four different WGDs within the legume family. The WGD that is shared by *Salix purpurea* and *Populus trichocarpa* is additionally modelled in all the analyses. Likelihood ratio tests (LRTs) were used to compare the most likely (nested) models with different numbers of WGDs. *P* values for the LRTs at different confidence levels are given in Rabier et al. (2013).

Gene Tree Reconcilation with Allopolyploidy

To visualize potential reticulation we have redrawn the filtered supernetwork (Whitfield et al., 2008) of Koenen et al. (submitted) with the Convex Hull method implemented in SplitsTree4 (Huson and Bryant, 2005). Potential branches in the species tree that could be involved in allopolyploidy for an analysis with GRAMPA (Gregg et al., 2017) were identified. Because the GRAMPA method is unable to infer multiple WGDs, we generated a filtered gene tree set that does not include duplications that stem from the previously identified independent events in Detarioideae and Papilionoideae so that these do not influence the reconciliation scores. To do this, we used the gene trees spanning the nitrogen-fixing clade generated for the WGDgc analysis and reduced Cercidoideae, Detarioideae and Papilionoideae to single

accessions (*Bauhinia tomentosa, Anthonotha fragrans* and *Medicago truncatula*, respectively), collapsing all duplications that are particular to these terminal taxa. An independent autopolyploidy event is not well established for Caesalpinioideae even though this subfamily showed a polyploid signal in Ks plots (Cannon et al., 2015). Therefore, we retained the four transcriptomes of A*lbizia julibrissin, Entada abyssinica, Inga spectabilis* and *Microlobius foetidus* since they were well-represented in gene trees. In this way we could test whether polyploidy in Caesalpinioideae is likely derived from independent autopolyploidy or allopolyploidy, or instead by an earlier WGD that is shared with other subfamilies. For the gene tree set used for this analysis, we first calculated average bootstrap scores for each tree, after which trees with <50% average support were excluded.

Fossil Time-calibration Priors

Fossils used to calibrate molecular clock analyses on the species tree, as described below, are listed in Table 1 and discussed here.

Non-legume Eudicot Fossils – These were taken from Magallón et al. (2015) and are thoroughly discussed in the supplementary information of that article. The numbers listed in Table 1 match those used in the Supplementary Information Methods 1 of Magallón et al. (2015). We have followed their fossil placements although our more limited taxon sampling means that some minimum ages are placed on deeper nodes. The only exception is the stem node of Fagales (calibration X14), which was here calibrated using the oldest fossil prior used by Xing et al. (2014). All minimum ages were updated to the latest version of the Geologic Time Scale (v. 4.0; Gradstein et al., 2012).

Legume Fossils – The selection of legume fossils used here for calibrating the divergence time estimation analyses differs from previous legume time tree studies (Lavin et al., 2005; Bruneau et al., 2008; Simon et al., 2009), both in the placement of fossils as well as in the minimum ages that some of these fossils represent. Calibrations

Table 1. Fossil calibrations used in the divergence time analyses.

Calibration[a]	Definition	Fossil	Age (Ma)
eudicots			
26	CG eudicots	Tricolpate pollen; England and Gabon[b]	126[c]
27	CG Ranunculales	*Teixeiraea lusitanica* – flower; Portugal[b]	113
38	CG Pentapetalae	Pentamerous flower with distinct calyx and corolla; USA[b]	100
48	SG Ericales	*Pentapetalum trifasciculandricus* – flowers; USA[b]	89.8
94	SG Myrtaceae	"Flower number 3" from the Table Nunatak Formation, Antarctica[b]	83.6
105	SG Brassicales	*Dressiantha bicarpelata* – flowers; USA[b]	89.8
112	CG Rosaceae	*Prunus wutuensis* – fruits; China[b]	49.4
116	SG Cannabaceae	*Aphananthe cretacea* and *Gironniera gonnensis* – fruits; Germany[b]	66
122	SG Juglandaceae	*Polyptera manningi* – fruits; USA[b]	64.4
133	SG *Populus*	*Populus wilmattae* – leaves, infructescences and fruits; USA[b]	37.8
X14	SG Fagales	*Protofagacea allonensis* – flowers; USA[d]	83.6
legumes			
A	SG Leguminosae	*Paracacioxylon frenguellii* – wood with vestured pits; Argentina[e]	63.5
C	SG *Cercis*	*Cercis parvifolia* – leaves and C. herbmeyeri – fruits; USA[f]	36
C[g]	SG *Bauhinia*	cf. *Bauhinia* – simple leaf with bilobed lamina; Tanzania[h]	46
F	SG Resin-producing clade	*Hymenaea mexicana* – vegetative and floral remains in amber; Mexico[i]	22.5

G	SG Detarioideae	*Aulacoxylon sparnacense* – wood and amber; France[j]	53
G[g]	SG Resin-producing clade	same as G	53
H[g]	CG Amherstieae	*Aphanocalyx singidaensis* – bifoliolate leaves; Tanzania[k]	46
I2	SG *Styphnolobium*/*Cladrastis*	*Styphnolobium* and *Cladrastis* – leaves and fruits; USA[l]	37.8
M2	SG Robinioid clade	*Robinia zirkelii* – wood; USA[m]	33.9
Q	SG Acacieae/Ingeae	Flattened polyads with 16 pollen grains; Brazil, Colombia, Cameroon and Egypt[n]	33.9
Q2	SG *Acacia* s.s.	Polyads with pseudocolpi; Australia[o]	23
Z	SG Caesalpinioideae	Bipinnate leaves; Colombia[p]	58

CG = Crown group; SG = Stem group; Ma = Million years ago.

[a] numbers 26, 27, 38, 48, 94, 105, 112, 116, 122 and 133 refer to calibrations from Magallón et al. (2015) as listed in their Supplementary Information Methods S1; letters A, D, F, G, I2, M2 and Q refer to calibrations from Bruneau et al. (2008) and/or Simon et al. (2009)
[b] Magallón et al. (2015) and references therein
[c] prior set as normal with standard deviation of 1.0, and truncated between minimum and maximum bounds of 113 and 136 Ma, respectively
[d] Xing et al. (2014) and reference therein
[e] Brea et al. (2008)
[f] Jia and Manchester (2014)
[g] alternative prior 1 as used in FLC analysis with 8 local clocks
[h] Jacobs and Herendeen (2004)
[i] Poinar and Brown (2002)
[j] De Franceschi and De Ploëg (2003)
[k] Herendeen and Jacobs (2000)
[l] Herendeen (1992)
[m] Lavin et al. (2003) and references therein
[n] Simon et al. (2009): Supplementary Information and references therein
[o] Miller et al. (2013)
[p] Wing et al. (2009)

Q2 and Z are used for the first time here. Calibrations A, D, F, G, I2, M2 and Q are labelled according to the schemes of Bruneau et al. (2008) and/or Simon et al. (2009), and differences from previous studies are discussed here. Other fossils used by Bruneau et al. (2008) and/or Simon et al. (2009) are not used here because of our sparser taxon sampling.

First, we did not fix the crown age of the family, which is critical as it is the most important node for which we want to estimate the age. The oldest definitive legume fossil, a fossil wood from the Early Paleocene of Patagonia (Brea et al., 2008), is used to set a minimum age on the stem node of the family at 63.5 Ma (calibration A, same node as in Bruneau et al. (2008) and Simon et al. (2009), but a new fossil and minimum age). This calibration is probably uninformative because of the long stem of the family, but it is included for completeness. The oldest crown group fossil, bipinnate leaves from the Late Paleocene of Colombia (Wing et al., 2009; Herrera et al., in press), is placed on the stem node of Caesalpinioideae with a minimum age of 58 Ma (calibration Z), a new calibration that has not been used in previous studies. This calibration renders the calibration of the stem of Papilionoideae (which is sister to Caesalpinioideae), with fossil flowers of *Barnebyanthus buchananensis* from the Paleocene-Eocene boundary at 56 Ma (Crepet and Herendeen, 1992), redundant.

We find the interpretation of some Early and Middle Eocene fossils, that were used in previous studies to calibrate lineages within crown group Cercidoideae and Detarioideae (Bruneau et al., 2008; Simon et al., 2009) to be problematic. Bruneau et al. (2008: Table 3) already pointed out the large discrepancy in age estimates of Detarioideae between calibrated and non-calibrated analyses. Given that this subfamily has a very long stem lineage, placing Early to Middle Eocene fossils within the crown group would require very high inferred substitution rates along the stem lineage, while at the same time implying a relatively low substitution rate for the Detarioideae crown group lineages (see Results). Cercidoideae are also subtended by a long stem lineage, leading to similar, although less severe substitution rate discrepancies than in

Detarioideae. We investigate and test this with molecular clock analyses with fixed local clocks, as described below. Here, we discuss the interpretation of these fossils as either stem or crown relatives and how we have calibrated lineages from subfamilies Cercidoideae and Detarioideae.

Bauhinia-like bilobed leaves from the Eocene of Tanzania (c. 46 Ma) (Jacobs and Herendeen, 2004) were used by Bruneau et al. (2008) and Simon et al. (2009) to calibrate the stem lineage of *Bauhinia* s.l.. This leaf type is highly characteristic for Cercidoideae and therefore the fossil is certainly representative of the subfamily. However, even though this type of leaf is not found in *Cercis*, which has been found to be sister to the rest of the genera in the subfamily (Bruneau et al., 2008; Wang et al., 2018), it may not provide a strong apomorphy for crown group Cercidoideae. Leaves in *Bauhinia* s.l. are variously bifoliolate, bilobed or entire, implying that entire leaves like those of *Cercis* have evolved multiple times independently, leading to homoplasy. This means that the bilobed leaves may have been present in the most recent common ancestor (MRCA) or stem relatives of Cercidoideae, and evolved to having an entire lamina in *Cercis.* If the Tanzanian fossils are a possible stem-relative of Cercidoideae, we consider the oldest definitive crown group fossil evidence to be the recently described *Cercis* fossil leaves and fruits from the Late Eocene of Oregon (Jia and Manchester, 2014), at c. 36 Ma (calibration C, a slightly older minimum age than used by Bruneau et al. (2008) and Simon et al. (2009)).

Bifoliolate leaves from the same fossil site in Tanzania as the *Bauhinia* fossil were ascribed to *Aphanocalyx* (Detarioideae) (Herendeen and Jacobs, 2000) based on distinctive venation patterns, after comparing the leaves to all extant legume genera with bifoliolate leaves. The fossil was used to calibrate the stem lineage of that genus by Bruneau et al. (2008) and Simon et al. (2009). The genus is deeply nested within Detarioideae, also meaning that the difference between age estimates from calibrated and uncalibrated analyses is large (46.0 vs 4.4 Ma; Bruneau et al., 2008: Table 3). While venation patterns can be diagnostic in many cases, they are often variable even

within modern genera and likely to be homoplasious. Therefore, these fossils might also represent an extinct lineage, possibly a stem relative of Detarioideae, that had evolved similar leaf morphology to extant *Aphanocalyx.* Moreover, the author of the most recent taxonomic account of *Aphanocalyx* (Wieringa, 1999), Jan Wieringa, does not accept this fossil as belonging to the genus. It also does not fit with the morphology-based phylogeny of *Aphanocalyx* which showed that bifoliolate leaves evolved recently and are derived within *Aphanocalyx* (Wieringa, 1999). In general, leaflet numbers are highly variable across Detarioideae, so relatives of fossils should not be sought only among other bifoliolate taxa.

Further evidence of Detarioideae from the Eocene is found at two localities within the Claiborne Formation in western Tennessee, USA. Fruits and leaflets from those sites are ascribed to the genus *Crudia* (Herendeen and Dilcher, 1990). As for the *Aphanocalyx* fossil, the affinities of the fossils were carefully evaluated before concluding that they are related to *Crudia.* Bruneau et al. (2008) and Simon et al. (2009) used this fossil to calibrate the stem of *Crudia* at 45 Ma, but as for the *Aphanocalyx* fossil age, an uncalibrated analysis finds a far younger age (6.9 Ma; Bruneau et al., 2008: Table 3). It is possible that in this case, an extinct detarioid lineage may have evolved morphological features similar to extant *Crudia* species independently. The raised venation on the fruit valves and twisted petiolules that most strongly resemble *Crudia*, for example, are both homoplasious across Detarioideae.

Fossil wood, flowers and amber of *Aulacoxylon sparnacense*, which has previously been interpreted as related to the extant genus *Daniellia* (Detarioideae), from the Early Eocene of the Paris basin (De Franceschi and De Ploëg, 2003), provide the most convincing evidence of fossils representing Early to Middle Eocene crown group members of Detarioideae. The fossil wood has vestured pits and resin canals like modern resin-producing Detarioideae, and the amber deposits are chemically similar to the Dominican ambers. Bruneau et al. (2008) considered the wood and flowers similar to *Daniellia*, but suggested they could also belong to a different genus of resin-

producing Detarieae. However, it is also possible that resin-production was already present in stem-relatives of Detarioideae. This is quite likely given that this trait is homoplasious across the resin-producing clade, having apparently been independently gained or lost several times, with only about half of the extant genera in the clade producing resin (Fougère-Danezan et al., 2007). If the production of resin evolved in the ancestral lineage of Detarioideae it would not require many more losses to account for the absence of the trait in the other lineages of the subfamily, because the resin-producing clade branches deeply within Detarioideae and the basal relationships of the subfamily are poorly resolved and understood (Bruneau et al., 2008; de la Estrella et al., 2018). Furthermore, the large majority of genera in the subfamily are confined to the large clade of Amherstieae, so perhaps only a single additional loss of the trait in the lineage leading to this clade could have produced this homoplasious pattern. This makes it possible that the Paris basin fossils belong to an extinct genus belonging to the stem group of Detarioideae. Therefore, the *Aulacoxylon* fossils can be used either to calibrate the stem node of the resin producing clade (calibration G^g, as done by Bruneau et al, (2008) and Simon et al., (2009) or the stem node of Detarioideae (calibration G), with a minimum age of 53 Ma .

For the disputed age of Dominican amber (Iturralde-Vinent and MacPhee, 1996), an intermediate age of 24 Ma was chosen by Bruneau et al. (2008), which was followed by Simon et al. (2009), but it is preferable to not consider an intermediate age as a valid minimum, but rather to use the minimum age that was estimated for Mexican amber that includes flowers of *Hymenaea mexicana*, the extinct species that presumably produced the amber (Poinar and Brown, 2002), and we calibrate the Detarieae s.s. stem node with a minimum age of 22.5 Ma (calibration F, a more inclusive node than in Bruneau et al, (2008) and Simon et al., (2009), and a different fossil age).

The calibration of the stem group of Styphnolobium and Cladrastis (calibration I2) is the same as used in Bruneau et al. (2008) and Simon et al. (2009), but the minimum age was updated to 37.8 Ma according to the latest version of the Geologic Time Scale

(v. 4.0; Gradstein et al., 2012), representing the end of the Middle Eocene (end of the Bartonian). Calibration M2 is the same as used in Simon et al., (2009) but since *Robinia* itself is not sampled here, we place it on the stem node of the robinioid clade (represented here by *Lotus japonicus*) and update the minimum age to the Eocene-Oligocene boundary at 33.9 Ma.

Bruneau et al. (2008) and Simon et al. (2009) also set the ages of several fossil calibrations at the midpoint of the Eocene, at 45 Ma. This led to a bias that was observable in an LTT plot of legumes (Koenen et al., 2013), and here we prefer to use the minimum boundary ages for these fossils. Although most of these calibrations are not used in our analyses due to sparser taxon sampling, we use one of these fossils, *Acacia*-like polyads, to calibrate the minimum stem age of the clade including all *Acacia* s.l. segregates at 33.9 Ma, the Eocene-Oligocene boundary (calibration Q, same node but younger age than Bruneau et al, (2008) and Simon et al., (2009)). Finally, we add calibration Q2, based on Australian Oligocene polyads with pseudocolpi (Miller et al., 2013), which suggest affinity with *Acacia* s.s., and we calibrate the stem node of that genus with a minimum age of 23 Ma, the Oligocene-Miocene boundary.

Divergence time analyses

Using SortaDate (Smith et al., 2018b), we analyzed the 1,103 gene trees from Koenen et al. (submitted) to estimate the total tree length (a proxy for sequence variation or informativeness), root-to-tip variance (a proxy for clock-likeness) and compatibility of bipartitions with the ML tree that was inferred using the full data set (the RAxML tree inferred with the LG4X model). We then selected the best genes for dating based on cutoff values that were arbitrarily chosen from the estimated values across gene trees: (1) total tree length greater than 5, (2) root-to-tip variance less than 0.005 and (3) at least 10% of the bipartitions in common with the ML tree. This yielded 36 genes, which were concatenated to have a total aligned length of 14462 amino acid

sites. We also used the ‘pxlstr’ program of the Phyx package (Brown et al., 2017) to calculate taxon-specific root-to-tip lengths from the ML tree, after pruning the Ranunculales, on which the tree was rooted. The values obtained were then used to define local clocks as described below. *Arabidopsis thaliana, Linum usitatissimum* and *Polygala lutea* were removed because of much higher root-to-tip lengths relative to their closest relatives. *Panax ginseng* was also removed because of a low root-to-tip length relative to the other sampled asterids, leaving a total of 72 taxa.

We used BEAST v.1.8.4 (Drummond et al., 2012) with various clock models to estimate divergence time estimates across the phylogeny based on the alignment of the selected 36 genes and the fossil calibrations described above. All analyses were run with the LG + G model of amino acid substitution and the birth-death tree prior, and using the ML tree to fix the topology. Fossil calibration priors were set as uniform priors between the minimum age as specified in Table 1 and a maximum age of 126 Ma (oldest fossil evidence of eudicots) as listed in Table S2, with the exception of the root node, for which we used a normal prior at 126 Ma with a standard deviation of 1.0 and truncated to minimum and maximum ages of 113 (the Aptian-Albian boundary) and 136 Ma (the oldest crown angiosperm fossil, see Magallón et al. (2015)). With these settings, we ran analyses under the uncorrelated lognormal (UCLN), strict (STRC), random (RLC) and 3 different fixed local (FLC) clock models. To specify the different FLC models, we examined root-to-tip length variation across subclades to specify biologically meaningful *a priori* clock partitions (Fig. S19). The 50kb-inversion clade of papilionoid legumes and the asterids (without *Panax ginseng*) have uniformly longer root-to-tip lengths than the remaining taxa across the tree and were each therefore assigned their own local clocks, with a different clock for all remaining taxa in the tree (this model referred to as FLC3, partitioning of taxa is illustrated in Supplementary Figure S19A). A more complex model was specified where the rosid rate was decoupled from the background rate and more clock partitions within the legumes were created for the mimosoids together with the *Cassia* clade because of their longer root-to-tip lengths

relative to other Caesalpinioideae and most of the rosid clade and also for the combined clade of Cercidoideae and Detarioideae. This more complex model is referred to as FLC6 (Fig. S19B). The most complex model (FLC8; Fig. S19C) was generated by further partitioning the combined clade of Cercidoideae and Detarioideae with a separate local clock for each subfamily, and one on their combined stem lineages (this most complex partitioning is also indicated with colored branches in Figures 6 and S16-17 and those of the other FLC models in Figures S14-15). The Ranunculales that were pruned for the root-to-tip length calculations were included in the background clock for each FLC model.

The separate clock partitions assigned to Cercidoideae and Detarioideae in the FLC8 model are particularly useful for evaluating the controversial placement of Early and Middle Eocene fossils within their crown groups. This was done by running two analyses under the FLC8 model, one with the same priors as the other analyses, and one with similar placements of these calibrations as in Bruneau et al. (2008) and Simon et al. (2009) (Table 1). Calibration C was replaced with a minimum age of 46 Ma on the stem of Bauhinia, based on fossil leaves from Tanzania (Herendeen and Jacobs, 2000; discussed in the previous paragraph). Calibration G^g was applied on the stem of the resin-producing clade (i.e. the crown node of Detarioideae) instead of on the stem node of Detarioideae. Calibration H^g is taken from Bruneau et al. (2008) and Simon et al. (2009), and is added in the alternative analysis to specify a minimum age of 46 Ma on the stem of *Anthonotha*, based on fossil leaves assigned to the closely related genus *Aphanocalyx* from Tanzania (Herendeen and Jacobs, 2000; but see previous paragraph). We refer to this calibration scheme as “alternative prior 1” (Table S2). Since a separate local clock is assigned to the combined stem lineages of Cercidoideae and Detarioideae, substitution rate estimates for stem and crown groups can be compared under both calibration schemes.

Maximum ages of fossil calibrations were set conservatively, and perhaps overly so, which can lead to a poorly formed joint marginal prior on node ages across the tree

(Phillips, 2015). Therefore, we also constructed an alternative prior with less conservative maxima as specified in Table S2 (“alternative prior 2”). These maxima represent boundary ages of older epochs from which the crown or stem group is not known, and in line with ages found by Magallón et al. (2015). These analyses serve to test the sensitivity of the UCLN model to the marginal prior.

Analyses sampling from the prior (without data) were run for 100 million generations, the strict clock and FLC3 and FLC6 analyses were run for 25 million generations and all other clock analyses were run for 50 million generations, and convergence was confirmed with Tracer v1.7.1 (Rambaut et al., 2018). For the non-prior analyses, the first 10% of the total number of generations was discarded as burn-in before summarizing median branch lengths and substitution rates with TreeAnnotator from the BEAST package.

To infer the ages of gene duplication nodes, we made four new subsets of gene trees for time-scaling. The first set includes all gene trees for which duplications were mapped on the collapsed legume backbone by Notung, but including only well-sampled taxa (see Table S1), and all other rosids as outgroup taxa. The other three sets were obtained by taking the sequences of all non-legume taxa in the nitrogen-fixing clade of angiosperms as outgroups alongside sequences of selected, well-sampled accessions for each of the subfamilies Caesalpinioideae, Detarioideae and Papilionoideae, creating separate sets of gene trees for each of these subfamilies together with the non-legume outgroup taxa. We chose these three subfamilies since they are well-sampled and their paleopolyploidy is well established. In this way we could assess whether the WGD events in different subfamilies occurred at different times or at the same time as would be expected if they are shared WGDs, although this in itself does not constitute evidence for shared events. For Detarioideae all four transcriptomes were included, for Caesalpinioideae only those of *Entada abyssinica*, *Microlobius foetidus*, *Albizia julibrissin* and *Inga spectabilis*, and for Papilionoideae the genomes of *Medicago truncatula*, *Glycine max*, *Phaseolus vulgaris* and *Arachis ipaensis*. For all of these sets,

sequences were realigned and new gene trees were inferred with RaxML, using the PROTGAMMAAUTO model. The resulting maximum likelihood trees were rooted with Notung with respect to the species tree relationships. For the family wide trees we further tested whether all legume sequences formed a monophyletic group to make sure no gene duplications pre-dating the divergence of the legumes were included. For each subfamily gene tree set we ran a phyparts analysis and all gene trees with duplications that mapped to the crown node of the subfamily were selected. All gene trees in the family-wide and subfamily specific sets were then individually time-scaled using penalized likelihood (Sanderson, 2002) as implemented in the R package ape (function 'chronos') (Paradis et al., 2004; Paradis, 2013). Based on simulations, it was shown that even although the correlated clock model estimates more accurate substitution rates, the strict clock estimates more accurate branch lengths (Paradis, 2013). Since it is our purpose to estimate ages and not rates, we therefore used the strict clock in these analyses, and set the smoothing parameter to 1 as done by Paradis (2013). The root age was set at 110 Ma for the family-wide gene tree set and to 105 Ma for the subfamily-specific gene tree sets based on the crown age estimates for the rosids and nitrogen-fixing clade of angiosperms from time-scaling analyses on the species tree (Figs. S6-S13). After time-calibration, ages of the duplication nodes were extracted and histograms and density plots of these were made in R.

RESULTS

After removing fragmentary sequences and gappy sites from the 9,282 homolog clusters generated by Koenen et al. (submitted), 640 clusters with large amounts of missing data were eliminated. From trees that were inferred from the remaining 8,642 homologs, we extracted different sets of rooted gene trees for analysis: (1) a set of 8,038 trees for the Phyparts analyses that include all taxa except Ranunculales that were used for rooting, (2) a set of 8,324 trees including only rosid taxa for the Notung and WGDgc analyses and (3) a set of 4,371 pruned trees with only taxa from the

nitrogen-fixing clade of angiosperms, including four Caesalpinioideae species and one species each of the remaining subfamilies, and average BS > 50%, for the GRAMPA analysis. Exemplar gene trees from the first set are shown in Figure S1. Because of the way these homolog sets were assembled, duplications that are restricted to a terminal lineage are not included, therefore testing for a WGD specific to Dialioideae or one within Cercidoideae (but excluding *Cercis*), both of which were suggested by Stai et al. (2019), is not possible with this data set. For time-calibrating the species tree, 36 informative and relatively clock-like genes were selected from the 1,103 orthologs of Koenen et al. (submitted). To estimate the timing of gene duplication nodes, we analysed 863 gene trees extracted from the Notung analysis including taxa from multiple subfamilies and 246, 250 and 272 trees including only Caesalpinioideae, Detarioideae and Papilionoideae, respectively. Table S2 gives an overview of which accessions were included per analysis, and the number of trees and sequences that were included per taxon. Alignments, gene trees and gene count data are included in Supplementary Data S1-S7.

Inferring Phylogenetic Locations of WGDs

In the Phyparts analysis, we find significantly elevated numbers of gene duplications at several nodes where WGDs are hypothesized to have occurred, including the previously documented *Salix/Populus* clade (Tuskan et al., 2006) and one subtending Pentapetalae, consistent with the known *gamma* hexaploidization associated with that clade (Jiao et al., 2012) (Figs 1a and S2). For the Pentapetalae clade, many homologs show more than one gene duplication at that node, given that the number of duplications (1,901) is nearly twice as high than the number of homologs with duplications (1,105), as expected for two consecutive rounds of WGD. Some of these duplications may also stem from older events, since missing data for the three non-Pentapetalae taxa in our dataset could mean that we do not find duplicates of older

WGDs in these taxa. In the legumes, high numbers of gene duplications at particular nodes suggest that there were three early WGD events, one at the base of the family, and one each subtending subfamilies Papilionoideae and Detarioideae (Figs 1a and S2). When applying a bootstrap filter to the homolog trees (≥50% bootstrap support), numbers of gene duplications are considerably lower, but the pattern is the same (Figs 1a and S2). At the root of the family, the number of gene duplications drops from 1,646 to 99 when applying this bootstrap filter, in line with the difficulty of resolving the deepest dichotomies of the legume phylogeny (Koenen et al., submitted). Notably, for the legume crown node we also find evidence for a significant fraction of homologs showing more than one gene duplication, with 1,646 duplications from only 1,229 homologs mapping to that node. This could suggest multiple rounds of WGD (e.g. Figs S1e and f), although some of these can be attributed to duplications in both paralog copies of genes duplicated at the Pentapetalae *gamma* event, while for many others support values across the tree are low. For other hypothesized WGDs, the numbers of homologs with more than one duplication at those nodes are much lower, suggesting they involved a single round of WGD.

Using gene tree reconciliation with Notung, we found similar results (Fig. 1b, S3 and S4), although in this case the Pentapetalae node was not included in the analysis. Furthermore, numbers of duplications particular to Detarioideae are higher than in the Phyparts analysis. The opposite is true for Papilionoideae, where instead Notung finds higher numbers of gene duplications on the node uniting Caesalpinioideae and Papilionoideae, and on several nodes within Papilionoideae. The differences between these two analyses are likely to be mainly attributed to the --rearrange method in Notung that was used to account for poor support in gene trees.

The likely placements of WGDs based on the phylogenetic locations of gene duplications were further tested with WGDgc, a probabilistic method based on gene count data harvested from the second, rosid gene tree set. The best scoring model with two WGDs has one WGD that is specific to Detarioideae and one that is shared by

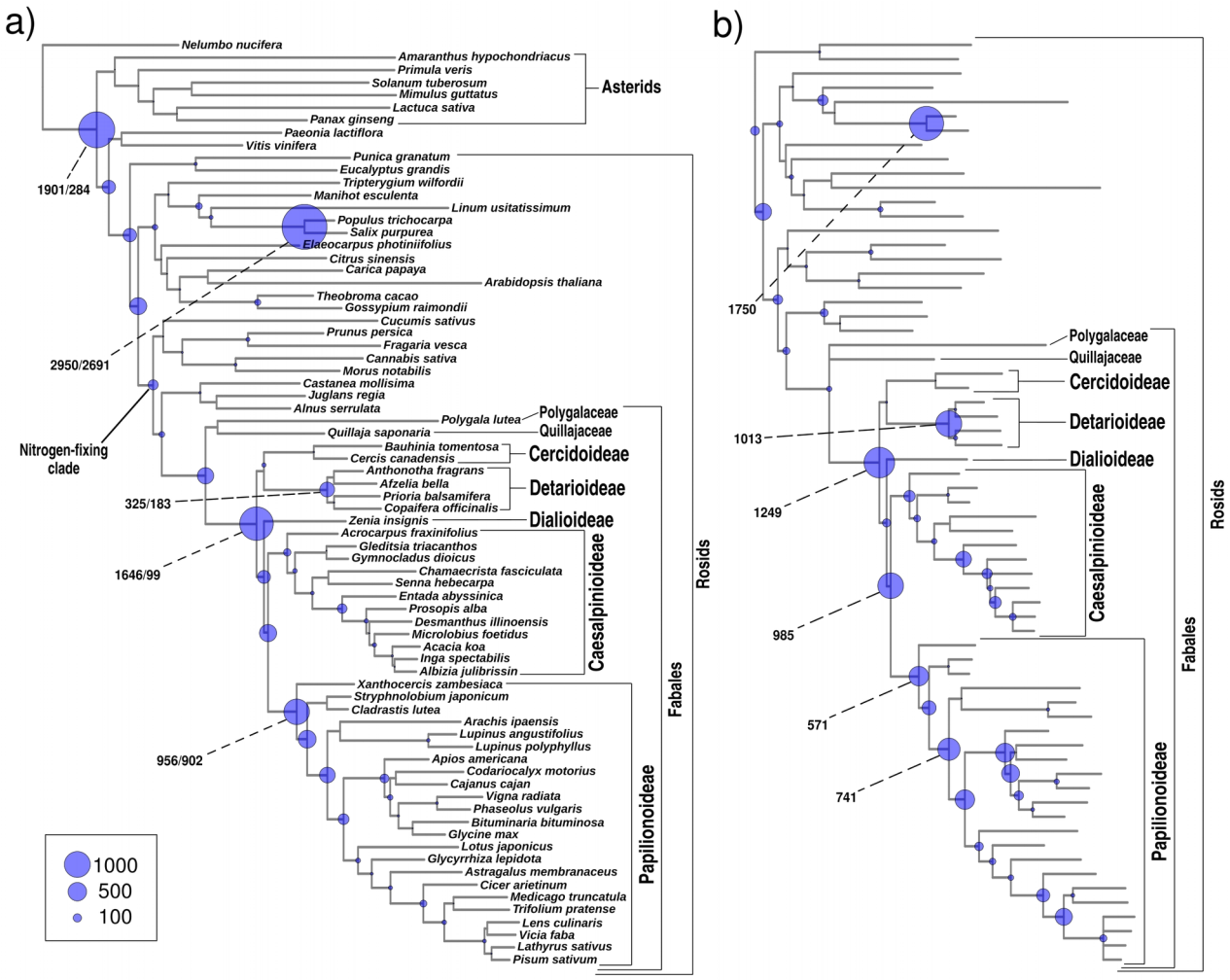


FIGURE 1. Numbers of gene duplications mapped over the species tree. a) Results from a phyparts analysis on the species tree topology of Koenen et al. (submitted) and b) results from a Notung analysis on the rosids portion of the same tree. Relative sizes of circles on nodes indicate the number of duplications as per the legend. Numbers of duplications are indicated for putative WGDs, in a) the two numbers are derived from ML topologies without and with a bootstrap filter of 50%, respectively.

Papilionoideae and Caesalpinioideae (Fig. 2a), which received a higher likelihood than a model with WGDs specific to Detarioideae and Papilionoideae (Fig. 2d), as well as to other models with two WGDs. When adding a third WGD specific to Papilionoideae, the LRT score of 25.76 suggests that this three-WGD model is significantly better at the α =

0.001 confidence level (*P* value > 9.550) (Fig. 2b). Other models with three WGDs received lower likelihood scores (Fig. 2e), where the second best scoring three-WGD model is that with independent WGDs in Caesalpinioideae, Detarioideae and Papilionoideae which corresponds to the results of Cannon et al. (2015) and Stai et al. (2019). Adding a fourth WGD on the legume crown node (Fig. 2c) further improves the likelihood, but the LRT score of 7.94 is only significant at a lower confidence level of α = 0.01 (*P* value > 5.412). Alternative placement of a fourth WGD within legumes (Fig. 2f) has a lower likelihood than placing it on the legume crown node and received an LRT score of 1.16 which is not significant even at a confidence level of α = 0.05 (*P* value > 2.706).

Distinguishing Between Auto- and Allopolyploidy Along the Legume Backbone

An allopolyploid event along the legume backbone could provide an alternative explanation for the high numbers of gene duplications mapping to the crown node of the legumes. Only one or a few of the subfamilies need to be derived from such an event in order for duplicate gene copies to map to the legume crown node if the parental lineages of the polyploid diverged from each other at the base of the family. Under this scenario no pan-legume WGD would need to be inferred and the subfamilies could each be derived from an independent WGD or be ancestrally non-polyploid as suggested by Cannon et al. (2015) and Stai et al. (2019), or a WGD could be shared across two or more subfamilies. In the filtered supernetwork, the Convex Hull method draws a complex tangle of ‘boxed’ relationships at the putative placements of WGDs as inferred with Phyparts, Notung and WGDgc: at the bases of the Papilionoideae, Detarioideae and the family as a whole (Fig. 3). This suggests that indeed at least three WGDs occurred early in the evolution of the legumes, one of which occurred along the backbone before or among the first divergences in the family. For most subfamilies, there is not much reticulation involving the root edge of each, except for

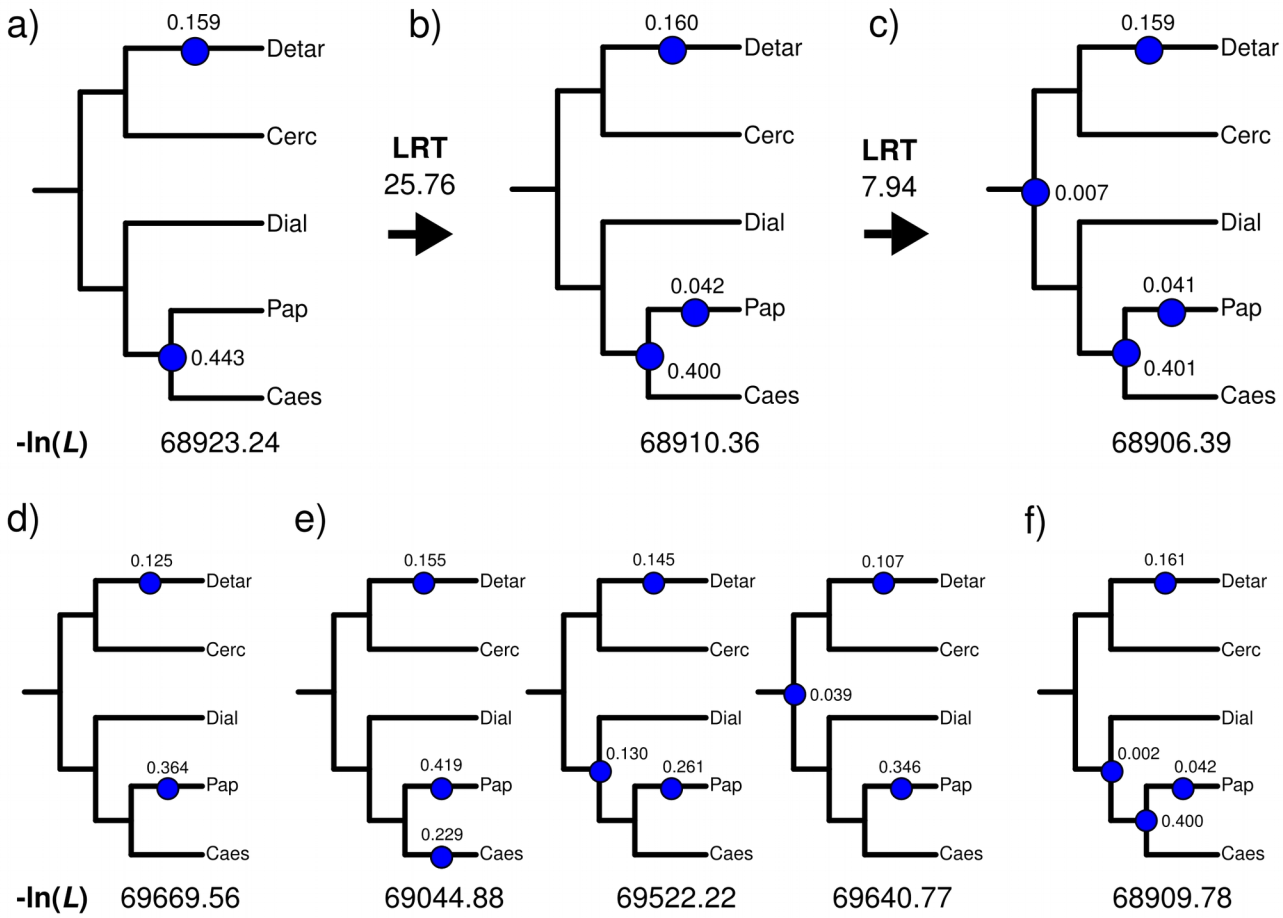


FIGURE 2. Possible placements of legume WGD events on the species tree, and their log-likelihoods based on the gene count method implemented in WGDgc. In the top row models with the highest likelihood scores are shown for a) two WGDs, b) three WGDs and c) four WGDs, with likelihood ratio test (LRT) scores indicated above the arrows between each panel. d) The second most likely model with two WGDs. e) The three next most likely models with three WGDs, from left to right: the model corresponding to results from Cannon et al. (2015) and Stai et al. (2019); an alternative model to b) with a shared WGD for Caesalpinioideae, Dialioideae and Papilionoideae; and the model with a pan-legume WGD as suggested by the Phyparts and Notung analyses (Fig. 1). f) The second most likely model with four WGDs. The WGD subtending *Populus* and *Salix* in the outgroup taxa is not shown but was included in all analyses. Caes = Caesalpinioideae, Cerc = Cercidoideae, Detar = Detarioideae, Dial = Dialioideae and Pap = Papilionoideae. Blue circles represent WGDs, the numbers above them indicate the estimated retention rates.

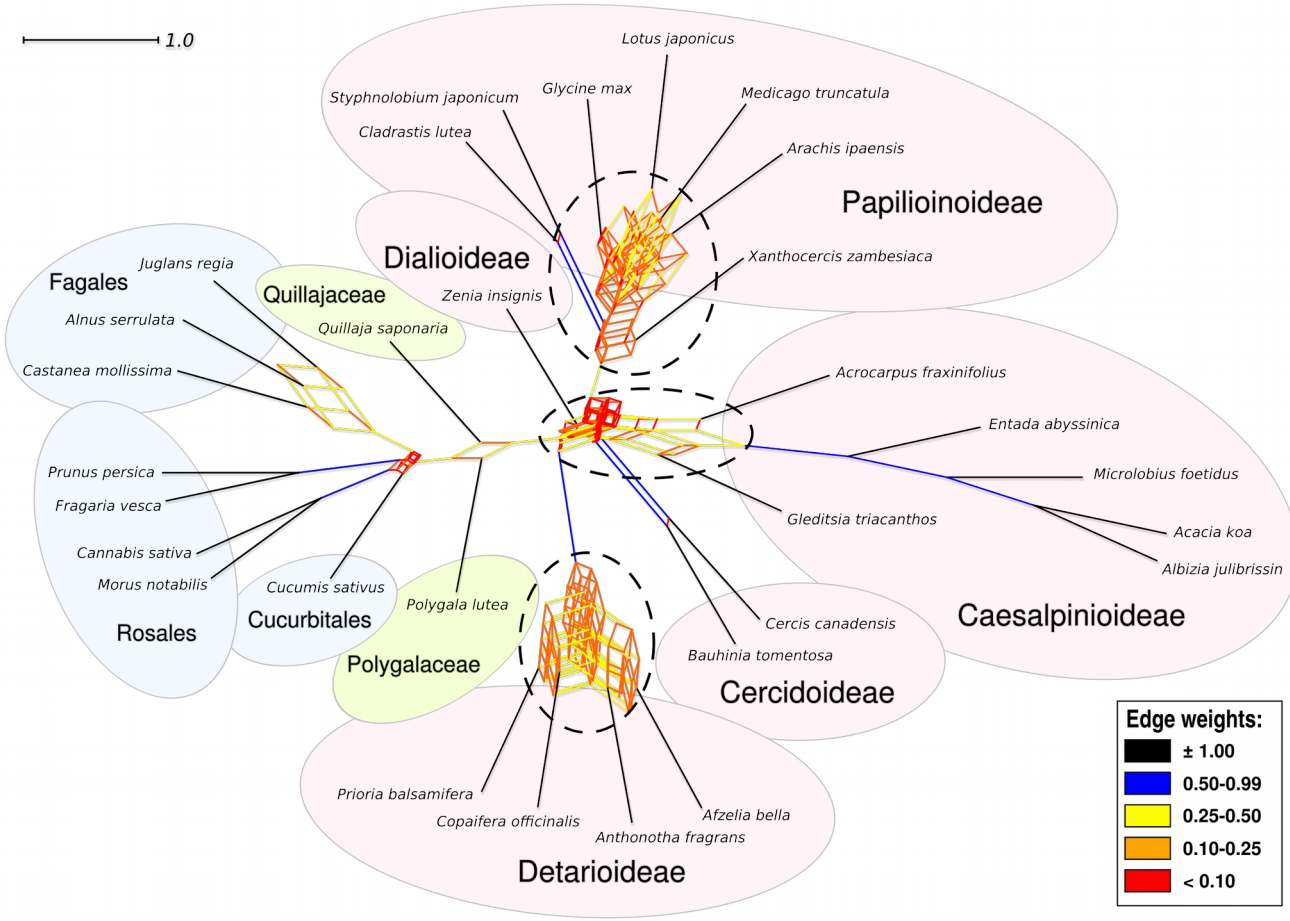


FIGURE 3. A filtered supernetwork drawn with the Convex Hull algorithm shows tangles of gene tree relationships at the bases of the legumes, and subfamilies Detarioideae and Papilionoideae, that correspond to WGDs, as well as possible reticulation at the base of Caesalpinioideae. The filtered supernetwork was inferred from the 1,103 1-to-1 ortholog gene tree set, and only bipartitions that received more than 80% bootstrap support in gene tree analyses were included. Edge lengths and colours are by their weight, a measure of prevalence of the bipartition that the edge represents among the gene trees. Ellipses with dashed outline indicate increased complexity at putative locations of WGDs.

Caesalpinioideae. This suggests that (at least) this subfamily could have resulted from an allopolyploid event.

The GRAMPA method identified eight multi-labeled (MUL) trees representing allopolyploid events (summarized in Fig. 4a-f), that had lower (better) reconciliation scores than the singly labeled species tree (Fig. 4g). MUL trees with just autopolyploidy

(Fig. 4h and i) received higher (worse) scores. The two best scoring MUL trees (Fig. 4a) included an allopolyploid event involving either Cercidoideae or Detarioideae as the second parental lineage for the clade combining the other three sampled subfamilies. The same second parental lineages are implied in the fourth and fifth best-scoring trees, for the Caesalpinioideae + Papilionoideae clade (Fig. 4c). Given that strong gene tree conflict was observed among the orthologs analysed by Koenen et al. (submitted), these MUL trees may receive better scores due to incomplete lineage sorting (ILS) and/ or gene tree estimation errors. The only high scoring MUL tree with an independent allopolyploid event restricted to Caesalpinioideae (Fig. 4f) scored only slightly better than the singly labeled tree. The remaining high scoring MUL trees involve a shared allopolyploidy event for Caesalpinioideae and Papilionoideae (Fig. 4b and e) or one in which it is shared with Dialioideae (Fig. 4d). The highest scoring of these involves an allopolyploid event from which are derived Caesalpinioideae and Papilionoideae with the second parental lineage stemming from a divergence that occurred before the first dichotomy in the species tree (Fig. 4b), in line with the high number of duplications mapped onto the legume crown node in Phyparts and Notung analyses (Fig. 1). Furthermore, an allopolyploid event shared between Caesalpininoideae and Papilionoideae is also in line with the high number of duplications mapping on the node that unites these two subfamilies in the Notung analysis (Fig. 1b).

Divergence Time Estimation

To establish whether the origin of the legumes is closely associated with the KPB, we performed clock dating in a Bayesian framework. Because the chloroplast phylogeny of Koenen et al. (submitted) shows large root-to-tip length variation, we refrained from using chloroplast data to infer divergence time estimates, and instead rely on the better suited nuclear data for this purpose, as suggested by Christin et al. (2014). We selected 36 informative and relatively clock-like nuclear genes and 20 fossil calibrations

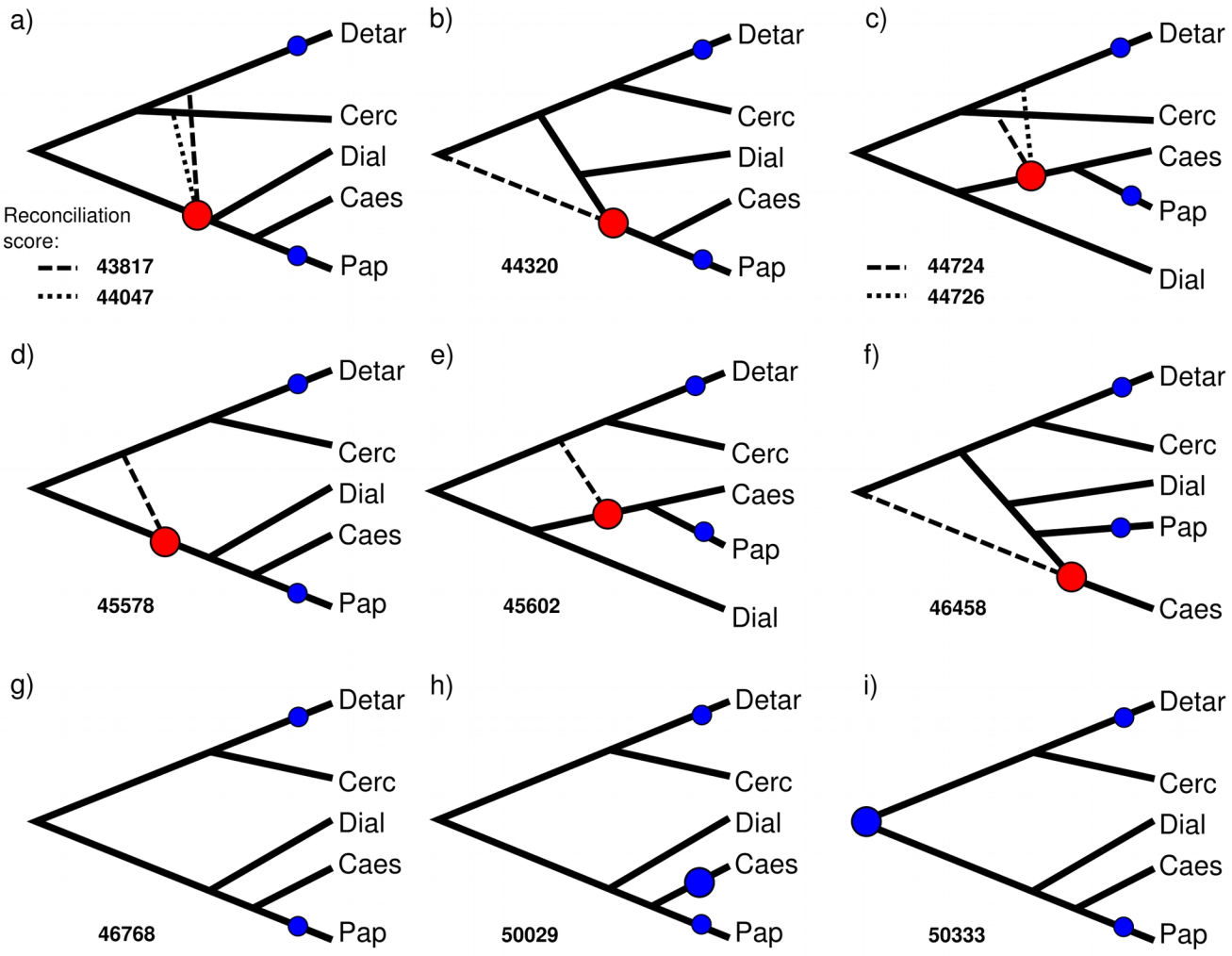
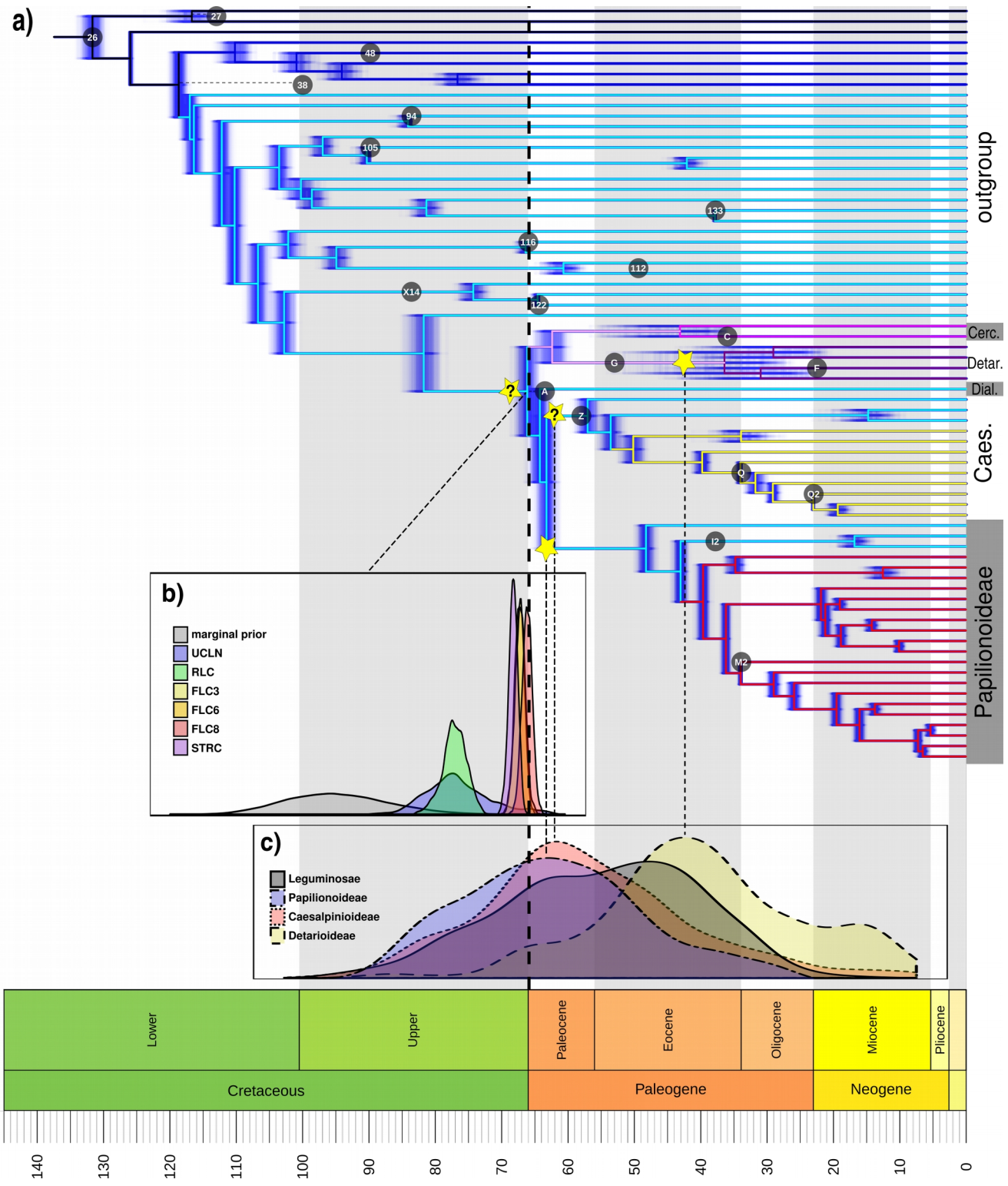


FIGURE 4. Different hypotheses involving allopolyploidy and their reconciliation scores in comparison to hypotheses involving only autopolyploidy. (a - f) All eight allopolyploid hypotheses that gave lower (better) reconcilation scores than (g), which represents the null hypothesis with no allopolyploidy. Hypotheses involving an additional autopolyploid event (h) in Caesalpinioideae or (i) at the legume crown node lead to higher (worse) reconciliation scores. Red circles indicate putative allopolyploidy events, large blue circles indicate putative autopolyploidy events taken into account in the analysis, small blue circles indicate autopolyploid events in Papilionoideae and Detarioideae that were not taken into account and removed from the input gene trees prior to the analysis. Solid lines represent the species tree topology; dashed lines connect to the putative second parental lineage. Caes = Caesalpinioideae, Cerc = Cercidoideae, Detar = Detarioideae, Dial = Dialioideae and Pap = Papilionoideae.

(Table 1). The oldest definitive fossil evidence of crown group legumes is from the Late Paleocene, consisting of bipinnate leaves from c. 58 Ma (Wing et al., 2009; Herrera et al., in press) and papilionoid-like flowers from c. 56 Ma (Crepet and Herendeen, 1992), representing Caesalpinioideae and Papilionoideae respectively. The older fossil woods with vestured pits, from the Early Paleocene of Patagonia (Brea et al., 2008) and the Middle Paleocene of Mali (Crawley, 1988), could represent stem relatives of the family (vestured pits are found in Papilionoideae, Caesalpinioideae and Detarioideae, so this is likely an ancestral legume trait). Based on this fossil evidence, c. 58 Ma can be considered the minimum age of the legume crown node. Molecular age estimates (95% HPD intervals) for the crown node range from 65.47-86.45 Ma and 73.46-81.18 Ma under the uncorrelated log-normal relaxed clock (UCLN) and the random local clock (RLC) models, respectively, to minima and maxima between 64.63 and 68.85 Ma under various fixed local clock (FLC) models (Table S3), the latter suggesting a close association of the origin of the legumes with the KPB (Fig. 5). Maximum clade credibility (MCC) trees for all clock analyses, with 95% HPD intervals indicated, are included in Supplementary Figures S6-S13, and 95% HPD intervals for nodes A-H are listed in Table S3.

Placement of Eocene fossils of Detarioideae and Cercidoideae within the crown groups of those clades (Bruneau et al., 2008; Simon et al., 2009; de la Estrella et al., 2017), yields older crown group estimates for these clades. However, with these calibrations (alternative prior 1 in Table S2), a more than 10-fold higher substitution rate along the stem lineages of these two subfamilies relative to the rates within both crown clades is inferred (c. 8.82×10^{-3} vs 0.69×10^{-3} substitutions per site per million years, with identical rates estimated independently for Cercidoideae and Detarioideae; Fig. S14a). This rate is also nearly five times higher than the mean rate across the tree as a whole (1.54×10^{-3} substitutions per site per million years), while the crown clades are estimated to have rates about half as high as the mean. Analyses with the same clock partitioning but calibrated with Late Eocene *Cercis* fossils and Mexican amber


a)
27
26
48
38
94
105
133
116
112
X14
122
C
G
F
?
A
?
Z
Q
Q2
I2
M2
outgroup
Cerc.
Detar.
Dial.
Caes.
Papilionoideae
b)
marginal prior
UCLN
RLC
FLC3
FLC6
FLC8
STRC
c)
Leguminosae
Papilionoideae
Caesalpinioideae
Detarioideae
Lower
Upper
Paleocene
Eocene
Oligocene
Miocene
Pliocene
Cretaceous
Paleogene
Neogene
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0

(PREVIOUS PAGE) FIGURE 5. The origin of the legumes is closely associated with the KPB. (a) Chronogram estimated with 8 fixed local clocks (FLC8 model) in BEAST, with the clock partitions indicated by colored branches, from an alignment of 36 genes selected as both clock-like and highly informative and hence well-suited for dating analyses. Blue shading represents 500 post-burnin trees ('densitree' plot) indicating posterior distributions of node ages. Yellow stars indicate putative legume WGD events. Labelled circles plotted across the phylogeny indicate placement and age of fossil calibrations listed in Table 1. (b) Prior and posterior distributions for the age of legumes under different clock models, as indicated in the legend. (c) Density plots of age estimates for duplication nodes in gene trees, for all duplications that mapped onto the legume crown node in the Notung analysis in grey and for duplications in the three well sampled subfamilies Papilionoideae, Caesalpinioideae and Detarioideae as indicated in the legend.

(*Hymenaea*) as the oldest crown group evidence for Cercidoideae and Detarioideae, respectively, do not infer such strong substitution rate shifts, with all clock partitions across the phylogeny estimated to have a substitution rate ranging from 0.96×10^{-3} to 2.53×10^{-3} substitutions per site per million years (Fig. S14b). Either way, different placements of these fossils have little influence on the crown age estimates for the family in the FLC analyses (Figs S11 and S12, Table 3).

Age estimates for duplication nodes show that (at least) Caesalpinioideae and Papilionoideae are derived from one or more WGDs that occurred close to the KPB (Fig. 5c and S15). The WGD specific to Detarioideae appears to have occurred more recently, in the Eocene. The duplication nodes corresponding to the legume backbone as inferred from the Notung analysis are likely a mixture of Detarioideae WGD duplications and older legume WGDs. This is surprising since it implies that detarioid paralogs do not always form sister clades in the gene trees, which could be caused by gene tree estimation errors or an allopolyploid origin for that subfamily. The large spread of ages found for the duplication nodes (Fig. 5c) may be attributed to substitution rate variation across genes, which, in the absence of fossil calibrations, is not accounted for. This may also have led to underestimation of ages for a proportion of gene tree duplication nodes given the long-tailed distributions towards the present (Fig. 5c and S15), which are also conflicting with crown age estimates for the subfamilies from the

dating analysis on the species tree (Fig. 5a). On the other hand, in the case of allopolyploidy, the estimated ages of duplication nodes reflect the divergence time of the two parental lineages rather than the allopolyploid event itself, leading to an overestimation of the timing of polyploidy.

DISCUSSION

We find evidence for at least three WGD events early in the evolution of the family, which further complicate the phylogenomic tangle characterised by lack of phylogenetic resolution and strong gene tree conflict at the base of the family documented by Koenen et al. (submitted). Time-calibration of the species tree suggests a close association of this complex origin of the legumes with the KPB. We discuss these findings and their relevance to understanding the early evolution of the third largest angiosperm family, the likely complications caused by WGDs on phylogenetic inferences in deep time, and the consequences of the KPB mass extinction event for plant evolution in the Cenozoic.

Locating WGD Events on the Phylogeny

Our analyses provide evidence for at least three WGD events early in the evolution of the legume family, one that occurred along either the stem lineage or the backbone of the family, plus independent WGDs subtending subfamilies Detarioideae and Papilionoideae. Our results suggest two likely hypotheses with regard to the oldest WGD event: (1) it is either placed on the stem lineage, representing a pan-legume WGD or (2) it involved allopolyploidy between two parental lineages that were derived from the first divergence within the family. The first hypothesis is supported by results from Phyparts and Notung analyses (Fig 1), while the WGDgc analysis only rejects a pan-legume WGD with the highest confidence interval in the LRT (Fig. 2). The second hypothesis is supported by the GRAMPA analysis (Fig. 4) and is more in line with the

idea that *Cercis* did not experience a legume-specific WGD, as suggested by Stai et al. (2019). Moreover, under the second hypothesis, duplicated genes would indeed also be reconciled onto the crown node of the family when using methods that do not account for allopolyploidy (Fig. 1). While this makes a pan-legume WGD less likely, all our results show that at least one WGD occurred among the first divergences of the family (Figs 1-4) and that this WGD is shared across more than one subfamily, rather than restricted to a single subfamily. Indeed, we show that it is unlikely that an independent WGD occurred in Caesalpinioideae (Figs 1 and 2), even in the case of allopolyploidy (Fig. 4). Most of the evidence instead suggests that Caesalpinioideae and Papilionoideae, perhaps together with Dialioideae, share a WGD (Figs 1b, 2 a, b and c, 4a-e), and that this was likely an allopolyploid event (Fig 4a-e). This would imply that subfamily Papilionoideae as a whole has undergone two rounds of WGD, which is overwhelmingly supported by the gene count method (Fig. 2b), with even some modest support for three rounds of WGD (Fig. 2c), but with lower confidence.

It is likely that missing data due to the inclusion of transcriptome data, rather than fully sampled genomes, has impacted some of our analyses. In particular, for Dialioideae, where only a single terminal is sampled, it remains uncertain whether that subfamily shares a WGD with Caesalpinioideae and Papilionoideae or not. The gene count method is likely to be particularly sensitive to missing data, as it does not take gene tree topology into account, thereby erroneously finding that a pan-legume WGD, if it did occur, was only shared by the better-sampled Caesalpinioideae and Papilionoideae for which high quality transcriptomes and genomes were used (Fig. 2a and b). Missing data could also impact the GRAMPA analysis with respect to identifying which parental lineages were involved in an ancient allopolyploid event and which subfamilies are derived from it. However, given that GRAMPA does take gene tree topology into account, the inference that allopolyploidy is more likely than autopolyploidy is likely to be robust, and moreover, none of the other results can reject allopolyploidy. Gene tree estimation errors and/or ILS will also have an influence on the GRAMPA

results and it is conceivable that this could erroneously lead to better reconciliation scores for allopolyploid hypotheses. More generally, different sources of gene tree discordance can impact on all of the analyses and indeed it appears difficult to distinguish among these (see below).

Until high quality assemblies and annotations of fully sequenced genomes are available for each of the subfamilies and closely related families, the impacts of missing data in these analyses are difficult to assess. Denser taxon sampling is also necessary to resolve the number and placement of WGDs with higher precision, accuracy and confidence. For Dialioideae, it will be highly desirable to include the genera *Poeppigia* and *Baudouinia* or *Eligmocarpus* to span the first two divergences of the subfamily (Zimmerman et al., 2017) and determine if a putative Dialioideae WGD was shared by all members of the subfamily. And for *Duparquetia orchidacea,* the sole member of Duparquetioideae, nuclear genomic and cytogenetic data are lacking, its phylogenetic placement within the family is based solely on chloroplast sequence data (Koenen et al., submitted) and any potential history of polyploidy for the taxon remains unknown.

Our results contrast with the conclusions of Cannon et al. (2015) and Stai et al. (2019) in that not all WGDs are restricted to individual subfamilies. The hypothesis of a pan-legume WGD contrasts most strongly with their hypothesis of four or five independent WGDs each confined to a single subfamily. However, an allopolyploid event shared across two or three subfamilies, but which excludes at least Cercidoideae and Detarioideae, also disagrees with their hypothesis, but is more in line with the idea that *Cercis* has not undergone a WGD since the origin of the legumes (Stai et al., 2019). The inferences of Cannon et al. (2015) and Stai et al. (2019) suffer from lack of rigorous hypothesis testing, which is especially problematic for the gene and chromosome count data used in Stai et al. (2019). While the evidence for an allopolyploid event within Cercidoideae appears to be relatively strong, the authors did not test whether an earlier pan-legume WGD can be rejected based on their gene count data (e.g. using the method of Rabier et al. (2013)). Similarly, the chromosome count data presented by Stai

et al. (2019) are not analysed based on chromosomal evolution modelling (Mayrose et al., 2009) across a robust species tree topology, and therefore provide rather weak evidence. Given the large variation in chromosome numbers observed across the legume family (LPWG, 2017; Stai et al., 2019) and across plant clades more generally (e.g. Hilu et al., 2004; Semple and Watanabe, 2009; Márquez-Corro et al., 2019), the assertion that *Cercis* would have retained the ancestral chromosome number throughout the whole of the Cenozoic should be tested for with formal analysis. Furthermore, the inference of Stai et al. (2019) that Cercidoideae is the sister-lineage to the rest of the legumes appears to stem from a phylogenetic analysis based on a single gene, the chloroplast gene *matK*, without evaluation of support for this relationship. In the densely sampled *matK* phylogeny of LPWG (2017) this relationship remained unresolved, while it now appears likely that Cercidoideae is the sister-lineage to Detarioideae rather than sister to the rest of the legumes (Koenen et al., submitted). Furthermore, the fact that *Cercis* is sister to the rest of Cercidoideae does not mean that it is more likely to have retained ancestral features than the remainder of the subfamily – the very concept of an "early-branching" lineage is a common misinterpretation of unbalanced phylogenies (Crisp and Cook, 2005). This is particularly relevant to the chromosome count results presented by Stai et al. (2019), where the haploid number of 7 in *Cercis* was suggested to be ancestral in legumes, but without analysing chromosomal evolution across the legume phylogeny. Haploid chromosome numbers of 6-8 are also found in Detarioideae, Caesalpinioideae and Papilionoideae, and are common in the latter, while paleopolyploidy in these three subfamilies is thought to be well established (Ren et al., 2019). Together with new data for Duparquetioideae and Dialioideae, stronger evidence for *Cercis* not having undergone a legume-specific WGD will potentially be able to reject a pan-legume WGD and lend further support to the allopolyploid hypothesis presented here, but based on the currently available evidence we believe this is premature.

Our divergence time analyses update previous analyses of Lavin et al. (2005), Bruneau et al. (2008) and Simon et al. (2009), and provide, to our knowledge, the first divergence time estimates for legumes based on nuclear genomic data. The age estimates under the FLC models and the strict clock model are mostly rather similar, but the RLC and UCLN models, that relax the clock assumption more, lead to older divergence time estimates. By allowing independent substitution rates on all branches, these models are potentially overfitting the data, to attempt to satisfy the marginal prior on node ages (Brown and Smith, 2017). As inferred from analyses run without data, the marginal prior that is constructed across all nodes of the tree, can be considered as "pseudo-data" (Brown and Smith, 2017), derived from the node calibration priors (based on fossil ages) and the branching process prior (constant birth-death model in our case), and should therefore not be overly informative on node ages. FLC and strict clock models lend greater weight to the molecular data and can overrule the marginal prior distributions on divergence times whilst still respecting hard maximum and minimum bounds of the fossil constraints on calibrated nodes, as suggested by our results. It is also clear from running analyses without data, that the marginal age prior on the (uncalibrated) crown node of the legumes is rather poorly informed, with the 95% HPD interval between 79.37-109.20 Ma (Fig. 5b), the minimum being much older than the oldest legume fossils, presumably caused by overly conservative maximum bounds on calibrated nodes (Phillips, 2015). UCLN and RLC analyses also inferred relatively high substitution rates for a few deep branches in the outgroup during the Lower Cretaceous, relative to the more derived and terminal branches of the tree (Figs S6 and S8), presumably to satisfy the poorly informed marginal priors. Phillips (2015) suggested that setting less conservative maxima on priors could remedy this problem, but our analysis with such prior settings shows little effect (Fig. S7), with some of the deepest branches still having much higher estimated substitution rates. Since there is

no evidence, nor any reason to assume, that substitution rates along those branches should be elevated relative to terminal branches, we conclude that this is indeed caused by overfitting of rate heterogeneity across branches under the influence of the marginal prior. Furthermore, the RLC analyses fitted c. 45 local clocks across the phylogeny, a rather high number relative to the total of 142 branches in the tree (implying a separate clock for every 3 branches), which is also indicative of overfitting. At the same time, this could be seen as evidence that the data are not the product of clock-like evolution, but it becomes difficult to estimate how much the clock deviates if the marginal prior on node ages is too influential. A more pragmatic approach is to use FLC analyses, by defining local clocks based on root-to-tip length distributions across clades and pruning outlier taxa (see Methods and Fig. S5). This approach accounts in large part for the violation of the molecular clock but it does not relax the clock to the extent that the marginal prior on node ages is given excessive weight relative to the molecular signal. Furthermore, because the genes we selected for divergence time estimation are reasonably clock-like and highly informative, it is desirable that these data inform the node ages with sufficient weight. One drawback of using this approach is that the relatively large amount of sequence data in combination with the FLC model results in estimates that appear unrealistically precise, and the discovery of new fossils may well prove the legumes to be slightly older. Nevertheless, the evidence presented here suggests that the legume crown age dates back to the Maastrichtian or Early Paleocene, likely within one or two million years before or after the KPB, although such high precision is not warranted due to the idiosyncrasies of the molecular clock.

Polyploidy (Senchina, et al., 2003) as well as the KPB itself (Berv and Field, 2018), have been implicated as potentially causing transient substitution rate increases, raising the possibility that substitution rates during the early evolution of the legumes could have deviated temporarily but markedly from the "background" rate of Cretaceous rosids. This would render the ages inferred for the first few dichotomies as well as those of the subfamilies less certain. The age estimates inferred for these nodes rely in large

part on the assumption that the substitution rate did not vary significantly within the different clock partitions, and most importantly within the rosid partition which includes most of the branches along the backbone of the family and the stem lineage subtending it. The WGD events that occurred along the legume backbone and within subfamilies Papilionoideae and Detarioideae could have affected substitution rates along those branches. By selecting for smaller stature and shorter generation times and reducing population sizes (Berv and Field, 2018), the KPB could additionally have resulted in increased rates along some or all of the stem lineages of the subfamilies, and, in the case of "hard" explosive diversification after the KPB, perhaps also along the legume stem lineage. A third factor that could influence node age estimates along the backbone of the family, is the gene tree incongruence observed for the first few legume divergences (Koenen et al., submitted), which is also observed among some of the 36 genes that were used for time-scaling. The divergence time analyses need to accommodate this incongruence within a single topology, meaning that additional substitutions need to be inferred for conflicting gene trees, which can inflate the branch lengths between rapid speciation events (Mendes and Hahn, 2016). Taken together, these three factors could mean that the time frame for the early evolution of the legumes appears inflated in our results, with (some of the) subfamily ages likely being slightly older than estimated here, as well as divergence of the subfamilies happening nearly simultaneously (hence the gene tree incongruence and lack of phylogenetic signal), rather than spanning the c. 3 - 5 million years inferred here (Figs 5 and S6-13). Potentially, even the legume crown age could be slightly older due to the effects of polyploidy, but not due to the KPB, because if the crown is older, the stem lineage would not have crossed the KPB.

Different interpretations of Eocene fossils of Cercidoideae and Detarioideae (see Material & Methods) lead to very different crown age estimates for these clades. As expected, this also leads to very different substitution rates along the stem lineages of these subfamilies, with rates increasing 10-fold when interpreting these fossils as crown

group members. While it cannot be ruled out that the stem lineages of Cercidoideae and Detarioideae experienced such markedly elevated substitution rates, it is unlikely that rates were five times higher relative to the rest of the eudicots across all 36 nuclear genes analysed, especially as these genes were chosen because of their approximately clock-like evolution, and given that these two clades comprise long-lived woody perennials. The idea that molecular information from extant taxa could inform that particular fossils are too old to belong to a crown clade is controversial. However, the test we have performed here is similar to the cross-validation method proposed by Near et al. (2005), which also uses molecular data to discover fossil calibration points that do not fit well with a larger set of fossils. Favouring those calibrations that do not lead to extreme substitution rate shifts is more parsimonious, and we believe that additional evidence is necessary to justify the inference of such a strong shift in substitution rates as that observed in the FLC8 analysis with alternative prior 1 (Fig. S12). While there seems little doubt that the Early Eocene fossils from the Mahenge in Tanzania and the Paris Basin in France do represent Cercidoideae and Detarioideae, the extreme substitution rate heterogeneity implied by their treatment as crown group members suggest that they may better be reinterpreted as stem-relatives of these subfamilies (see additional discussion about the affinities of these fossils in Material & Methods).

While we are not able at this point to confidently distinguish between a “hard” or “soft explosive” model of early diversification of the legumes, it is clear that the early radiations of the legume subfamilies all occurred in the Cenozoic. While stem age estimates of each subfamily are remarkably close to each other, crown age estimates are strikingly different (but see the discussion above on potential effects of polyploidy and the KPB on substitution rates and ages of subfamilies). Caesalpinioideae are found to have the oldest crown age (late Paleocene), followed by Papilionoideae with a crown age in the Early Eocene. Both of these subfamilies therefore likely diversified considerably during the PETM and Eocene climatic optimum, when tropical forests extended far into the Northern Hemisphere and paratropical forests occurred on the

coast of Antarctica. This is in line with the numerous legume fossil taxa known from the Eocene of North America, often of uncertain affinities, but with a majority ascribed to Caesalpinioideae and Papilionoideae (Herendeen, 1992). There is also fossil evidence of Early and Middle Eocene stem-relatives of Cercidoideae and Detarioideae (as discussed above and in Methods), but their crown group divergences are most likely placed in the Late Eocene or Oligocene. Our results suggest extinction of stem-relatives of these two subfamilies, most likely related to Late Eocene and Oligocene cooling, and subsequent diversification of the crown groups during the Oligocene and Miocene, when both groups become diverse at several fossil sites (e.g. Wang et al., 2014; Lin et al., 2015; Poinar, 1991; Poinar and Brown, 2002). Although it remains uncertain whether the crown group divergence of Detarioideae occurred in the (Late) Eocene or the Oligocene, the younger age of the subfamily inferred here contrasts with previous views of the evolutionary trajectories of this subfamily dating back into the Paleocene, comprising relatively slowly evolving lineages (de la Estrella et al., 2017), and with Amazonian subclades within Detarioideae conforming to the museum model of tropical rainforest diversification (Schley et al., 2018). This has important implications for our understanding of the origins of tropical African plant diversity, since Detarioideae dominate the canopy of many equatorial African rainforests, as well as being an important group in African savannas (de la Estrella et al., 2017). Our results for Detarioideae suggest that the extant diversity in tropical Africa, in particular the large diversity in tribe Amherstieae, is of relatively recent origin following a major turnover event at the Eocene-Oligocene boundary, which also affected other plant groups such as palms (Pan et al., 2006). This more recent diversification of detarioids is also more in line with the widely proposed recent assembly of the savanna biome (Cerling et al., 1997; Bouchenak-Khelladi et al., 2009; Maurin et al., 2014).

The Impact of the KPB on Plant Diversification

The impacts of the KPB mass extinction event on plant diversity are the focus of debate, with several studies claiming that extinction was less severe for plants than across marine and terrestrial faunas (Nicholls and Johnson, 2008; Cascales-Miñana and Cleal, 2014; Silvestro et al., 2015). However, our results suggest that the massive KPB turnover event likely played a critical role in the evolution of plant taxa. Our analyses indicate that the origin of crown group legumes is closely associated with the KPB. The analyses employing FLCs even suggest that potentially only a single legume ancestor crossed the KPB to give rise to the six main lineages during the early Paleocene, conforming to a “hard explosive” model. However, across the different analyses, part of the posterior density of the crown age estimate falls in the late Maastrichtian (Fig. 5), suggesting a “soft explosive” model, with the six main lineages diverging in the Late Cretaceous and crossing the KPB, giving rise to the crown groups of the modern subfamilies in the Cenozoic. These different explosive models have been used to describe the origin and early diversification of the placental mammals, although other studies have lent support to “short fuse” or “long fuse” models (summarized in Phillips, 2015: Fig. 1). For birds, the timing of diversification relative to the KPB has also been controversial (Ksepka and Phillips, 2015), but it now appears likely that the Neoaves underwent explosive radiation from a single ancestor that crossed the KPB (Suh, 2016). Apart from Placentalia and Neoaves, recent studies on frogs (Feng et al., 2017) and fishes (Alfaro et al., 2018) have also demonstrated rapid diversification following the KPB, suggesting this is a common pattern across many terrestrial and marine animal groups. We present here, to our knowledge, the first example of a major plant family whose origin and initial diversification appears to be closely linked to the KPB. This is notable because a recent family-level paleobotanical study suggested that the KPB did not constitute a mass extinction event for plants (Cascales-Miñana and Cleal, 2014). Phylogenetic studies in some plant families originating in the Cretaceous

also lack any evidence of a significant effect of the KPB on diversification (e.g. Annonaceae (Couvreur et al., 2011a) and Arecaceae (Couvreur et al., 2011b)), except for the smaller plant family Menispermaceae (Wang et al., 2012), which shows increased diversification following the KPB. In contrast, fern diversification appears to have been strongly affected, with some groups of ferns showing much reduced diversity in the Cenozoic compared to earlier times (Lehtonen et al., 2017), and especially epiphytic groups of ferns showing increased diversification rates since the KPB (Schuettpelz and Pryer, 2009). Furthermore, the generic-level study of Silvestro et al. (2015) showed high extinction rates for non-flowering plant groups during the late Cretaceous, and elevated origination rates for angiosperms during the Paleocene, in line with the pattern we observe for the origins of legume diversity. Thus, even if extinction was less severe for plants than for animals at the KPB, the Paleocene was nevertheless a time of major origination of lineages across biota, and we expect further examples of KPB-related accelerated plant diversification to be discovered when inferring larger angiosperm timetrees.

The Added Complications of Paleopolyploidy on Evolutionary Inferences in Deep Time

The recent proliferation of genomic data is revealing just how prevalent repeated WGDs have been in the history of the angiosperms (e.g. Wendel, 2015; Soltis et al., 2016; Yang et al., 2018; Cai et al., 2019; Conover et al., 2019) and how many large angiosperm clades are characterized by genome triplications (e.g. Pentapetalae, Brassicaceae, Asteraceae, Solanaceae). Here we show that there were also multiple WGDs during the early history of the legumes. It has been suggested that angiosperm WGDs are non-randomly distributed through time and significantly clustered around the KPB (Fawcett et al., 2009; Vanneste et al., 2014; Lohaus and Van de Peer, 2016), and we show that two of the early legume WGDs are also temporally close to the KPB (Fig. 5), lending further support to the idea that polyploid survival and establishment were

enhanced at or soon after the KPB with its associated rapid turnover of lineages (Lohaus and Van de Peer, 2016; Levin and Soltis, 2018). WGDs have also been hypothesized to trigger accelerated rates of lineage diversification at least in some lineages, albeit potentially after a time lag (Schranz et al., 2012; Tank et al., 2015; Landis et al., 2018; Smith et al., 2018a). The three legume WGDs we detected are each associated with rapid divergence of lineages as indicated by lack of support and short internodes along the legume backbone (LPWG, 2017; Koenen et al., submitted) and at the bases of subfamilies Detarioideae (de la Estrella et al., 2017 and 2018) and Papilionoideae (Cardoso et al., 2012 and 2013). Polyploidy could have helped ancestral legumes and other plant lineages to both survive the mass extinction event and rapidly diversify owing to differential gene loss and other processes of diploidization (Adams and Wendel, 2005; Dodsworth et al., 2016). Increased polyploid speciation and reduced diploid speciation in the wake of the KPB (Levin and Soltis, 2018) would then lead to over-representation of these WGD-derived lineages in the extant flora and clustering of WGDs around the KPB. On the other hand, many paleopolyploidy events that significantly pre- and post-date the KPB are known (e.g. Angiospermae (Jiao et al., 2011), Pentapetalae (Jiao et al., 2012), Salicaceae (Tuskan et al., 2006), Caryophyllales (Yang et al., 2018), *Gossypium* (Wendel, 2015)), Malpighiales (Cai et al., 2019), including in legumes (e.g. *Glycine*, Genisteae, the *Leucaena* group, *Vachellia)*, and more extensive sampling of recently diversified groups may well reveal a weaker pattern of clustering around the KPB, or a pattern of WGDs associated with episodes of rapid global change more generally (Cai et al., 2019).

Alongside rapid diversification and consequent lack of phylogenetic signal, the WGD events along the legume backbone and subtending subfamilies Detarioideae and Papilionoideae are likely to have contributed to the difficulties of obtaining phylogenetic resolution for the deep nodes in these clades (Cardoso et al., 2012 and 2013; de la Estrella 2018; Koenen et al., submitted). WGDs may have promoted increased lineage diversification rates resulting in short internodes and ILS. If the polyploidy event

happened some time before the first divergences in the legume family, or in the case of allopolyploidy, this could have led to divergent gene copies prior to lineage splitting which should make orthology detection easier. However, if the polyploidy event happened shortly before rapid cladogenesis, potentially a large fraction of paralogous gene copies would not have diverged at this point, making orthology detection challenging. In both cases, paralogous or homoeologous gene copies will have subsequently been differentially lost, pseudogenized or sub- or neo-functionalized, further complicating correct orthology detection (Wendel, 2015; Cheng F. et al., 2018). Together with ILS, this could explain the large fraction of gene trees supporting alternative topologies at the root of the legumes (Koenen et al., submitted). An allopolyploid event involving two or more early legume lineages (Fig. 4) offers an alternative explanation for gene tree discordance, but discriminating between these alternatives is not straightforward. It is notable that several other large plant clades, such as Pentapetalae (Zeng et al., 2017), Asteraceae (Barker et al., 2016; Huang et al., 2016), Brassicaceae (Couvreur et al., 2010; Huang et al., 2015) and Malvaceae (Conover et al., 2019), also appear to show similar lack of resolution in clades subtended by WGDs to that revealed here for the legume family and subfamilies Papilionoideae and Detarioideae. This suggests that the association of polyploidy with rapid divergence, which leads to a lack of phylogenetic signal and gene tree conflict, is potentially a common feature in the evolution of angiosperms and the origination of major plant clades.

A large number of homolog clusters do not show gene duplications along the backbone of the legumes or any of the subfamilies, suggesting that loss of paralog copies is widespread, as observed for ancient WGDs more generally (Adams and Wendel, 2005; Dehal and Boore, 2005; Brunet et al., 2006; Scannel et al., 2007). If many of those losses occurred along the stem lineages of the six subfamilies after their divergence, this could lead to different paralog copies being retained in different lineages, adding to conflict among gene trees. Loss of paralog copies along stem

lineages of subfamilies will also make it difficult to distinguish whether a gene duplication corresponds to a WGD that is shared among two or more subfamilies, or whether it represents a subfamily-specific nested WGD. Lack of support in those homolog trees showing gene duplications further complicates this issue, making it potentially extremely challenging to accurately reconstruct the history of WGDs and phylogenetic relationships. Given these difficulties, sampling a wider range of complete genomes will be important, since with transcriptome data it is unknown whether duplicate gene copies are lost or simply not expressed in the tissue from which the RNA was extracted. Furthermore, increased taxon sampling will help to counteract negative impacts of missing data, since particular duplicate gene copies may have been lost in all species sampled here, but not necessarily across the whole clade or subfamily which those species represent. Despite all these complications, the set of analyses presented here make it possible to reject some hypotheses such as an independent WGD subtending Caesalpinioideae, while also guiding us to formulate a new hypothesis involving ancient allopolyploidy (Fig. 4b representing the most likely scenario) that can potentially reconcile the large number of gene duplications inferred to have occurred at the root of the legumes (Fig. 1), with the presumed non-polyploid history of *Cercis* (Stai et al., 2019). In combination with ILS, this allopolyploid scenario can explain the discordant gene tree topologies observed at the base of the legume phylogeny (Koenen et al., submitted) and the distribution of duplicate gene copies across subfamilies, highlighting the complexity of the initial diversification and early genome evolution of the legumes.

Moreover, the observed complexity may be only an approximation of the full complexity of genome evolution and polyploidy that occurred in legumes in association with the KPB. These WGD events occurred c. 66 Ma and much of the evidence has been obscured by subsequent genome reorganization and loss of the large majority of duplicate gene copies. These issues limit the level of complexity that can be inferred for such ancient events compared to more recently evolved polyploidy. For instance, many

cases of polyploid complexes are known in the angiosperms, in which within a group of closely related species, recurrent allo- and autopolyploidy have led to extremely complex genomic relationships and variable ploidy levels, such as in the well-studied perennial soybean polyploid complex (Doyle, 2004). If a similar polyploid complex gave rise to the six major legume lineages, these could have had different ploidy levels with differing ancestries of subgenomes in cases of allopolyploidy.

Finally, alongside the need to further develop and extend methods to analyze ancient WGDs, we stress again the need for further genome sequencing in legumes and other Fabales, something that will be forthcoming as part of the 10KP initiative (S. Cheng S. et al., 2018). Currently, high quality well annotated genome assemblies are only available for Papilionoideae. Having similar data for all six legume subfamilies and closely related families would make it possible to further disentangle the early genome evolution of legumes by comparing conserved synteny blocks, detecting genomic rearrangements and reconstructing chromosomal evolution and the ancestral karyotype, as has recently been done for vertebrates (Sacerdot et al., 2018) and birds (Damas et al., 2018), as well as providing ample other opportunities to further enhance our understanding of legume evolution and diversification.

Concluding Remarks

It is becoming increasingly clear that the origin and early evolution of the legumes followed a complex scenario with multiple nested auto- and/or allopolyploidy events, and rapid divergence of the six main lineages against the background of a mass extinction event that led to major turnover in the Earth's biota and biomes. WGD likely contributed to the survival and evolutionary diversification of the legumes in the wake of the KPB mass extinction event, and contributed to the rise to ecological dominance of legumes in early Cenozoic tropical forests. At the same time, these events make it more difficult to reconstruct aspects of the early evolutionary history of the clade, including

evolutionary relationships, divergence time estimates and the phylogenetic location of the WGD events themselves. The similarities between the origins of the legumes and those of other major Cenozoic clades such as mammals and birds are striking. All three of these prominent Cenozoic clades show recalcitrant basal polytomies and parallel trajectories of rapid early divergence closely associated with the KPB, further emphasizing the importance of the KPB mass extinction event and the earth system succession that followed in its aftermath (Hull, 2015) in shaping the modern biota.

FUNDING

This work was supported by the Swiss National Science Foundation (Grant 31003A_135522 to C.E.H.); the Department of Systematic & Evolutionary Botany, University of Zurich; the Natural Sciences and Engineering Research Council of Canada (Grant to A.B.), the U.K. National Environment Research Council (Grant NE/I027797/1 to R.T.P.), and the Fonds de la Recherche Scientifique of Belgium (Grant J.0292.17 to O.H.).

ACKNOWLEDGEMENTS

We thank the S3IT of the University of Zurich for the use of the ScienceCloud computational infrastructure and the Functional Genomics Center Zurich (FGCZ) for library preparation and sequencing, and Robin van Velzen, Steven Cannon and an anonymous reviewer for constructive feedback that greatly improved the manuscript.

AUTHOR CONTRIBUTIONS

EK and CH designed the study, EK performed the analyses and prepared the manuscript, PH assisted with interpretation of fossil data and all authors contributed to data collection and writing of the final version of the manuscript.

REFERENCES

Adams K.L., Wendel J.F. 2005. Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8(2):135–141.

Alfaro M.E., Faircloth B.C., Harrington R.C., Sorenson L., Friedman M., Thacker C.E., Oliveros C.H., Černý D., Near T.J. 2018. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nat. Ecol. Evol. 2:688–696.

Barker M.S., Li Z., Kidder T.I., Reardon C.R., Lai Z., Oliveira L.O., Scascitelli M., Rieseberg L.H. 2016. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am. J. Bot. 103:1203–1211.

Barreda V.D., Cúneo N.R., Wilf P., Currano E.D., Scasso R.A., Brinkhuis H. 2012. Cretaceous/Paleogene Floral Turnover in Patagonia: Drop in Diversity, Low Extinction, and a Classopollis Spike. PLoS ONE 7(12):e52455.

Berv J.S., Field D.J. 2017. Genomic signature of an avian Lilliput Effect across the K-Pg extinction. Syst. Biol. 67(1):1–13.

Bouchenak-Khelladi Y., Anthony Verboom G., Hodkinson T.R., Salamin N., Francois O., Chonghaile G.N., Savolainen V. 2009. The origins and diversification of C4 grasses and savanna-adapted ungulates. Glob. Change Biol. 15(10):2397–2417.

Brea M., Zamuner A.B., Matheos S.D., Iglesias A., Zucol A.F. 2008. Fossil wood of the Mimosoideae from the early Paleocene of Patagonia, Argentina. Alcheringa. 32:427–441.

Brown J.W., Smith S.A. 2017. The past sure is tense: on interpreting phylogenetic divergence time estimates. Syst. Biol. 67:340–353.

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx: phylogenetic tools for unix. Bioinformatics. 33:1886–1888.

Bruneau A., Mercure M., Lewis G.P., Herendeen P.S. 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. Botany. 86:697–718.

Brunet F.G., Crollius H.R., Paris M., Aury J.M., Gibert P., Jaillon O., Laudet V., Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. 23(9):1808–1816.

Cai, L., Xi, Z., Amorim, A.M., Sugumaran, M., Rest, J.S., Liu, L. and Davis, C.C., 2019. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. New Phytol. 221: 565-576.

Cannon S.B., Sterc L., Rombauts S., Sato S., Cheung F., Gouzy J., Wang X., Mudge J., Vasdewani J., Schiex T., Spannagl M. 2006. Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes Proc. Natl. Acad. Sci. USA. 103:14959–14964.

Cannon S.B., McKain M.R., Harkess A., Nelson M.N., Dash S., Deyholos M.K., Peng Y., Joyce B., Stewart Jr C.N., Rolf M., Kutchan T. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol. Biol. Evol. 32(1):193–210.

Cardoso D., de Queiroz L.P., Pennington R.T., de Lima H.C., Fonty E., Wojciechowski M.F., Lavin M. 2012. Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. Am. J. Bot. 99:1991–2013.

Cardoso D., Pennington R.T., de Queiroz L.P., Boatwright J.S., Van Wyk B.-E., Wojciechowski M.F., Lavin M. 2013. Reconstructing the deep-branching relationships of the papilionoid legumes. S. Afr. J. Bot. 89:58–75.

Cascales-Miñana B., Cleal C.J. 2014. The plant fossil record reflects just two great extinction events. Terra Nova. 26:195–200.

Cerling T.E., Harris J.M., Macfadden BJ.., Leakey M.G., Quade J., Eisenmann V., Ehleringer J.R. 1997. Global vegetation change through the Miocene/Pliocene boundary. Nature. 389:153–158.

Cheng F., Wu J., Cai X., Liang, J., Freeling M., Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploidy plants. Nat. Plants 4: 258-268.

Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.-M., Li F.-W., Melkonian B., Mavrodiev E.V., Sun W., Fu Y., Yang H., Soltis D.E., Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K.-S. 2018. 10KP: a phylodiverse genome sequencing plan. GigaScience. 7:1–9.

Christin P.-A., Spriggs E., Osborne C.P., Strömberg C.A.E., Salamin N., Edwards E.J. 2014. Molecular dating, evolutionary rates, and the age of the grasses. Syst. Biol. 63:153–165.

Claramunt S., Cracraft J. 2015. A new time tree reveals Earth history's imprint on the evolution of modern birds. Sci. Adv. 1(11):e1501005.

Conover J.L., Karimi N., Stenz N., Ané C., Grover C.E. Skema, C., Tate J.A., Wolff K., Logan S.A., Wendel J.F., Baum D.A., 2019. A Malvaceae mystery: A mallow maelstrom of genome multiplications and maybe misleading methods? J Integrative Pl. Biol. 61: 12-31.

Cooper A., Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary Boundary: molecular evidence. Science. 275:1109–1113.

Couvreur T.L.P., Franzke A., Al-Shehbaz I.A., Bakker F.T., Koch M.A., Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. 27:55–71.

Couvreur T.L.P., Pirie M.D., Chatrou L.W., Saunders R.M.K., Su Y.C.F., Richardson J.E., Erkens R.H.J. 2011a. Early evolutionary history of the flowering plant family Annonaceae: steady diversification and boreotropical geodispersal. J. Biogeogr. 38:664–680.

Couvreur T.L.P., Forest F., Baker W.J. 2011b. Origin and global diversification patterns of tropical rain forests: inferences from a complete genus-level phylogeny of palms. BMC Biol. 9:44.

Crawley M. 1988. Palaeocene wood from the Republic of Mali. Bull. Br. Mus. (Nat. Hist.) Geol. 44:3–14.

Crepet W.L., Herendeen PS. 1992. Papilionoid flowers from the early Eocene of southeastern North America. In: Herendeen P.S., Dilcher D.L., editors, Advances in legume systematics part 4: The fossil record. Richmond, UK: Royal Botanic Gardens, Kew. p. 43–55.

Criscuolo A., Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10:210.

Crisp M.D., Cook L.G. 2005. Do early branching lineages signify ancestral traits? Trends Ecol. Evol. 20: 122-128.

Damas J., Kim J., Farré M., Griffin D.K., Larkin D.M. 2018. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. Genome Biol. 19(1):155.

De Franceschi D., De Ploëg G. 2003. Origine de l'ambre des faciès sparnaciens (Éocène inférieur) du Bassin de Paris: le bois de l'ambre producteur. Geodiversitas. 25:633–647.

de la Estrella M., Forest F., Wieringa J.J., Fougère-Danezan M., Bruneau A. 2017. Insights on the evolutionary origin of Detarioideae, a clade of ecologically dominant tropical African trees. New Phytol. 214(4):1722–1735.

de la Estrella M., Forest F., Klitgård B., Lewis G.P., Mackinder B.A., de Queiroz L.P., Bruneau A. 2018. A new phylogeny-based tribal classification of subfamily Detarioideae, an early branching clade of florally diverse tropical arborescent legumes. Sci. Rep. 8(1):6884.

Dehal P., Boore J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3(10):e314.

Dodsworth S, Chase M.W., Leitch A.R. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? Bot. J. Linn. Soc. 180(1):1–5.

dos Reis M., Donoghue P.C.J., Yang Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol. Lett. 10:20131003.

Doyle J.J., Doyle J.L., Rauscher J.T. and Brown A.H.D. 2004. Evolution of the perennial soybean polyploid complex (Glycine subgenus Glycine): a study of contrasts. Biol. J. Linn. Soc. 82(4):583-597.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Fawcett J.A., Maere S., Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous – Tertiary extinction event. Proc. Natl. Acad. Sci. USA. 106:5737–5742.

Feng Y.-J., Blackburn D.C., Liang D., Hillis D.M., Wake D.B., Cannatella D.C., Zhang P. 2017. Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. Proc. Natl. Acad. Sci. USA. 114(29):E5864–E5870.

Fougère-Danezan M., Maumont S., Bruneau A. 2007. Relationships among resin-producing Detarieae sl (Leguminosae) as inferred by molecular data. Syst. Bot. 32(4):748–761.

Gradstein F.M., Ogg J.G., Schmitz M.D., Ogg G.M. 2012. The Geologic Time Scale 2012. Boston, USA: Elsevier.

Gregg W.T., Ather S.H. and Hahn M.W. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. 66(6):1007-1018.

Herendeen P.S., Dilcher D.L. 1990. Reproductive and vegetative evidence for the occurrence of Crudia (Leguminosae, Caesalpinioideae) in the Eocene of southeastern North America. Bot. Gaz. 151:402–413.

Herendeen P.S., Dilcher D.L. 1992. Advances in legume systematics part 4. The fossil record. Richmond, UK: Royal Botanic Gardens, Kew..

Herendeen P.S. 1992. The fossil history of the Leguminosae from the Eocene of southeastern North America. In: Herendeen P.S., Dilcher D.L., editors, Advances in legume systematics part 4. The fossil record. Richmond, UK: Royal Botanic Gardens, Kew. pp. 85–160.

Herendeen P.S., Jacobs B.F. 2000. Fossil legumes from the Middle Eocene (46.0 Ma) Mahenge Flora of Singida, Tanzania. Am. J. Bot. 87:1358–1366.

Herrera F., Carvalho M.R., Wing S.L., Jaramillo C., Herendeen P.S. 2019. Middle to Late Paleocene Leguminosae fruits and leaves from Colombia. Austr. Syst. Bot. In press.

Hilu K.W. 2004. Phylogenetics and chromosomal evolution in the Poaceae (grasses). Aust. J. Bot. 52(1):13-22.

Huang C.-H., Sun R., Hu Y., Zeng L., Zhang N., Cai L., Zhang Q., Koch M.A., Al-Shehbaz I., Edger P.P., Pires J.C., Tan D.-Y., Zhong Y., Ma H. 2015. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. 33:394–412.

Huang C.-H., Zang C., Liu M., Hu Y., Gao T., Qi J., Ma H. 2016. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. Mol. Biol. Evol. 33:2820–2835.

Hull P. 2015. Life in the aftermath of mass extinctions. Curr. Biol. 25:R941–R952.

Huson D.H. and Bryant D. 2005. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23(2):254-267.

Iturralde-Vinent M.A., MacPhee R.D.E. 1996. Age and paleogeographical origin of Dominican amber. Science. 273:1850–1852.

Jacobs B.F., Herendeen P.S. 2004. Eocene dry climate and woodland vegetation in tropical Africa reconstructed from fossil leaves from northern Tanzania. Palaeogeogr. Palaeocl. 213:115–123.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C., Nabholz B., Howard J.T., Suh A. 2014. Whole genome analyses resolve the early branches in the tree of life of modern birds. Science. 346:1320–1331.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature. 491(7424):444–448.

Jia H., Manchester S.R. 2014. Fossil Leaves and Fruits of Cercis L. (Leguminosae) from the Eocene of Western North America. Int. J. Plant Sci. 175:601–612.

Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature. 473:97–100.

Jiao Y., Leebens-Mack J., Ayyampalayam S., Bowers J.E., McKain M.R., McNeal J., Rolf M., Ruzicka D.R., Wafula E., Wickett N.J., Wu X., Zhang Y., Wang J., Zhang Y., Carpenter E.J., Deyholos M.K., Kutchan T.M., Chanderbali A.S., Soltis P.S., Stevenson D.W., McCombie R., Pires J.C., Wong G.K.-S., Soltis D.E., DePamphilis C.W. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 13(1):R3.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30(4):772–780.

Keller G. 2014. Deccan volcanism, the Chicxulub impact, and the end-Cretaceous mass extinction: Coincidence? Cause and effect?, in Keller G., and Kerr A.C., eds., Volcanism, Impacts, and Mass Extinctions: Causes and Effects. Geol. S. Am. S. 505:57–89.

Koenen E.J.M., De Vos J.M., Atchison G.W., Simon M.F., Schrire B.D., De Souza E.R., de Queiroz L.P., Hughes C.E. 2013. Exploring the tempo of species diversification in legumes. S. Afr. J. Bot. 89:19–30.

Koenen E.J.M., Ojeda D.I., Steeves R., Migliore J., Bakker F.T., Wieringa J.J., Kidner C., Hardy O.J., Pennington R.T., Bruneau A., Hughes C.E. Submitted. Large-

scale genomic sequence data support a near-simultaneous evolutionary origin of all six legume subfamilies. New Phytologist.

Ksepka D.T., Phillips M.J. 2015. Avian diversification patterns across the K-Pg boundary: influence of calibrations, datasets, and model misspecification. Ann. Mo. Bot. Gard. 100(4):300–328.

Landis J.B., Soltis D.E., Li Z., Marx H.E., Barker M.S., Tank D.C., Soltis P.S. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. 105(3):348–363.

Lavin M., Wojciechowski M.F., Gasson P., Hughes C., Wheeler E. 2003. Phylogeny of robinioid legumes (Fabaceae) revisited: Coursetia and Gliricidia recircumscribed, and a biogeographical appraisal of the Caribbean endemics. Syst. Bot. 28:387–409.

Lavin M., Herendeen P.S., Wojciechowski M.F. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. Syst. Biol. 54:575–594.

Lehtonen S., Silvestro D., Karger D.N., Scotese C., Tuomisto H., Kessler M., Peña C., Wahlberg N., Antonelli A. 2017. Environmentally driven extinction and opportunistic origination explain fern diversification patterns. Sci. Rep. 7(1):4831.

Levin D.A., Soltis D.E. 2018. Factors promoting polyploid persistence and diversification and limiting diploid speciation during the K–Pg interlude. Curr. Opin. Plant Biol. 42:1–7.

Lin Y., Wong W.O., Shi G., Shen S., Li Z. 2015. Bilobate leaves of Bauhinia (Leguminosae, Caesalpinioideae, Cercideae) from the middle Miocene of Fujian Province, southeastern China and their biogeographic implications. BMC Evol. Biol. 15:252.

Lohaus R., Van de Peer Y. 2016. Of dups and dinos: evolution at the K/Pg boundary. Curr. Opin. Plant Biol. 30:62–69.

LPWG (Legume Phylogeny Working Group). 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. Taxon. 66:44–77.

Magallón S., Gómez-Acevedo S., Sánchez-Reyes L.L., Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 207:437–453.

Márquez-Corro J.I., Martín-Bravo S., Spalink D., Luceño M. and Escudero M. 2019. Inferring hypothesis-based transitions in clade-specific models of chromosome number evolution in sedges (Cyperaceae). Molecular phylogenetics and evolution.

Maurin O., Davies T.J., Burrows J.E., Daru B.H., Yessoufou K., Muasya A.M., Bank M., Bond W.J. 2014. Savanna fire and the origins of the ‘underground forests’ of Africa. New Phytol. 204(1):201–214.

Mayrose I., Barker M.S. and Otto S.P. 2009. Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst. Biol. 59(2):132-144.

McElwain J.C., Punyasena S.W. 2007. Mass extinction events and the plant fossil record. Trends Ecol. Evol. 22:548–557.

McKey D. 1994. Legumes and nitrogen: The evolutionary ecology of a nitrogen-demanding lifestyle. In: Sprent J.I., McKey D., editors. Advances in legume systematics part 5. The nitrogen factor. Richmond, UK: Royal Botanic Gardens, Kew. p. 211–228.

Mendes F.K., Hahn M.W. 2016. Gene tree discordance causes apparent substitution rate variation. Syst. Biol. 65(4):711–721.

Meredith R.W., Janecka J.E., Gatesy J., Ryder O.A., Fisher C.A., Teeling E.C., Goodbla A., Eizirik E., Simão TL., Stadler T., Rabosky D.L. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science. 334(6055):521–524.

Miller J.T., Murphy D.J., Ho S.Y.W., Cantrill D.J., Seigler D. 2013. Comparative dating of Acacia: combining fossils and multiple phylogenies to infer ages of clades with poor fossil records. Aust. J. Bot. 61:436–445.

Mudge J., Cannon S.B., Kalo P., Oldroyd G.E.D., Roe B.A., Town C.D and Young N.D. 2005. Highly syntenic regions in the genomes of soybean, Medicago truncatula, and Arabidopsis thaliana. BMC Plant Biol. 5:15.

Near T.J., Meylan P.A., Shaffer H.B., Meyer A.E.A. 2005. Assessing concordance of fossil calibration points in molecular clock studies: An Example Using Turtles. Am. Nat. 165(2):137–146.

Nicholls D.J., Johnson K.R. 2008. Plants and the K-T boundary. Cambridge, UK: Cambridge University Press.

O'leary M.A., Bloch J.I., Flynn J.J., Gaudin T.J., Giallombardo A., Giannini N.P., Goldberg S.L., Kraatz B.P., Luo Z.X., Meng J., Ni X. 2013. The placental mammal ancestor and the post–K-Pg radiation of placentals. Science 339(6120):662–667.

Pan A.D., Jacobs B.F., Dransfield J., Baker WJ. 2006. The fossil history of palms (Arecaceae) in Africa and new records from the Late Oligocene (28–27 Mya) of north-western Ethiopia. Bot. J. Linn. Soc. 151(1):69–81.

Paradis E., Claude J. and Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289-290.

Paradis, E., 2013. Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. Mol. Phylogenet. Evol. 67(2):436-444.

Phillips M.J. 2015. Geomolecular dating and the origin of placental mammals. Syst. Biol. 65(3):546–557.

Phillips M.J., Fruciano C. 2018. The soft explosive model of placental mammal evolution. BMC Evol. Biol. 18:104.

Poinar Jr G.O. 1991. Hymenaea protera sp. n. (Leguminoseae, Caesalpininoideae) from Dominican amber has African affinities. Experientia 47:1075–1082.

Poinar Jr G.O., Brown A.E. 2002. Hymenaea mexicana sp. nov. (Leguminosae: Caesalpinioideae) from Mexican amber indicates Old World connections. Bot. J. Linn. Soc. 139:125–132.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Rabier C.E., Ta T. and Ané C. 2013. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. Mol. Biol. Evol. 31(3):750-762.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67(5):901–904.

Ren L., Huang W., Cannon S.B. 2019. Reconstruction of ancestral genome reveales chromosome evolution history for selected legume species. New Phytol.

Ruprecht C., Lohaus R., Vanneste K., Mutwil M., Nikoloski Z., Van de Peer Y., Persson S. 2017. Revisiting ancestral polyploidy in plants. Sci. Adv. 3(7):e1603195.

Sacerdot C., Louis A., Bon C., Berthelot C., Crollius H.R. 2018. Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. 19(1):166.

Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19(1):101-109.

Scannell D.R., Frank A.C., Conant G.C., Byrne K.P., Woolfit M., Wolfe K.H. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc. Natl. Acad. Sci. USA. 104(20):8397–8402.

Schley R.J., de la Estrella M., Pérez-Escobar O.A., Bruneau A., Barraclough T., Forest F., Klitgård B. 2018. Is Amazonia a ‘museum’ for Neotropical trees? The evolution of the Brownea clade (Detarioideae, Leguminosae). Mol. Phylogenet. Evol. 126:279–292.

Schranz M.E., Mohammadin S., Edger P.P. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. 15:147–153.

Schuettpelz E., Pryer K.M. 2009. Evidence for a Cenozoic radiation of ferns in an angiosperm-dominated canopy. Proc. Natl. Acad. Sci. USA. 106(27):11200–11205.

Semple J.C. and Watanabe K., 2009. A review of chromosome numbers in Asteraceae with hypotheses on chromosomal base number evolution. In Funk et al. (eds). Systematics, evolution and biogeography of Compositae. IAPT, Vienna, pp. 61-72.

Senchina D.S., Alvarez I., Cronn R.C., Liu B., Rong J., Noyes R.D., Paterson A.H., Wing R.A., Wilkins T.A., Wendel J.F. 2003. Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol. Biol. Evol. 20(4):633–643.

Silvestro D., Cascales-Miñana B., Bacon C.D., Antonelli A. 2015. Revisiting the origin and diversification of vascular plants through a comprehensive Bayesian analysis of the fossil record. New Phytol. 207(2):425–436.

Simon M.F., Grether R., de Queiroz L.P., Skema C., Pennington R.T., Hughes C.E. 2009. Recent assembly of the Cerrado, a Neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. Proc. Natl. Acad. Sci. USA. 106:20359–20364.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:150.

Smith S.A., Brown J.W., Yang Y., Bruenn R., Drummond C.P., Brockington S.F., Walker J.F., Last N., Douglas N.A., Moore M.J. 2018a. Disparity, diversity, and duplications in the Caryophyllales. New Phytol. 217(2):836–854.

Smith S.A., Brown J.W., Walker J.F. 2018b. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PLoS One. 13(5):e0197433.

Soltis D.E., Visger C.J., Marchant D.B., Soltis P.S. 2016. Polyploidy: pitfalls and paths to a paradigm. Am. J. Bot. 103:1146–1166.

Springer M.S., Meredith R.W., Teeling E.C., Murphy W.J. 2013. Technical comment on “The placental mammal ancestor and the post–K-Pg radiation of placentals”. Science. 341:613.

Stai J.S., Yadav A., Sinou C., Bruneau A., Doyle J.J., Fernández-Baca D. and Cannon S.B. 2019. Cercis: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. Front. Plant Sci. 10:345.

Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Stolzer M., Lai H., Xu M., Sathaye D., Vernot B. and Durand D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28(18):i409-i415.

Suh A., Smeds L., Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13(8):e1002224.

Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. Zool. Scr. 45:50–62.

Tank D.C., Eastman J.M., Pennell M.W., Soltis P.S., Soltis D.E., Hinchliff C.E., Brown J.W., Sessa E.B., Harmon L.J. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytol. 207:454–467.

Teeling E.C., Hedges S.B. 2013. Making the impossible possible: rooting the tree of placental mammals. Mol. Biol. Evol. 30:1999–2000.

Tuskan G.A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N., Ralph S., Rombauts S., Salamov A., Schein J. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 313(5793):1596–1604.

Vajda V., Raine J.I., Hollis C.J. 2001. Indication of global deforestation at the Creataceous-Tertiary boundary by New Zealand fern spike. Science. 294:1700–1702.

Vajda V., Bercovici A. 2014. The global vegetation pattern across the Cretaceous–Paleogene mass extinction interval: A template for other extinction events. Global and Planet. Change. 122:29–49.

Vanneste K., Baele G., Maere S., Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. Genome Res. 24(8):1334–1347.

Wang H., Moore M.J., Soltis P.S., Bell C.D., Brockington S.F., Alexandre R., Davis C.C., Latvis M., Manchester S.R., Soltis D.E. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc. Natl. Acad. Sci. USA. 106(10):3853–3858.

Wang W., Ortiz R.D.C., Jacques F.M.B., Xiang X.-G., Li H.-L., Lin L., Li R.-Q., Liu Y., Soltis P.S., Soltis D.E., Chen Z.-D. 2012. Menispermaceae and the diversification of tropical rainforests near the Cretaceous–Paleogene boundary. New Phytol. 195:470–478.

Wang Q., Song Z., Chen Y., Shen S., Li Z. 2014. Leaves and fruits of Bauhinia (Leguminosae, Caesalpinioideae, Cercideae) from the Oligocene Ningming Formation of Guangxi, South China and their biogeographic implications. BMC Evol. Biol. 14:88.

Wang Y.-H., Wicke S., Wang H., Jin J.-J., Chen S.-Y., Zhang S.-D., Li D.-Z., Yi T.-S. 2018. Plastid genome evolution in the early-diverging legume subfamily Cercidoideae (Fabaceae). Front. Plant Sci. 9:138.

Wendel J.F. 2015. The wondrous cycles of polyploidy in plants. Am. J. Bot. 102:1753–1756.

Whitfield J., Cameron S.A., Huson D., Steel M. 2008. Filtered Z-closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees. Syst. Biol. 57:939–947.

Wieringa J.J. 1999. Monopetalanthus exit. A systematic study of Aphanocalyx, Bikinia, Icuria, Michelsonia and Tetraberlinia (Leguminosae, Caesalpinioideae). Agric. Univ. Wagening. Pap. 99(4):1–320.

Wilf P., Johnson K.R. 2004. Land plant extinction at the end of the Cretaceous: a quantitative analysis of the North Dakota megafloral record. Paleobiology. 30:347–368.

Wing S.L., Herrera F., Jaramillo C.A., Gómez-Navarro C., Wilf P., Labandeira C.C. 2009. Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest. Proc. Natl. Acad. Sci. USA. 106:18627–18632.

Xing Y.X., Onstein R.E., Carter R.J., Stadler T., Linder H.P. 2014. Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity and turnover rates are disconnected. Evolution. 68:2821–2832.

Yang Y., Smith S.A. 2014. Orthology inference in nonmodel nrganisms using transcriptomes and low-loverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events within Caryophyllales, including two allopolyploidy events. New Phytol. 217:855–870.

Zeng L., Zhang N., Zhang Q., Endress P.K., Huang J and Ma H. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytol. 214:1338–1354.

Zimmerman E., Herendeen P.S., Lewis G.P. and Bruneau A., 2017. Floral evolution and phylogeny of the Dialioideae, a diverse subfamily of tropical legumes. Am. J. Bot. 104(7):1019-1041.

Zwaenepoel A., Van de Peer Y. 2019. Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates. Mol. Biol. Evol.

Supplementary Information (see Appendix IV on page 242)

Table S1. Taxon occupancy per analysis..

Table S2. Age intervals specified for the fossil calibration priors under different alternative priors.

Table S3. Node age estimates and priors (95% HPD intervals) of nodes A-H in the different analyses.

Figure S1. Examples of homolog clusters with gene duplications in legumes that passed the bootstrap filter. Yellow stars behind nodes indicate locations of gene duplications, numbers on nodes indicate bootstrap support. The plotted gene trees are extracted from (a) cluster3675_1rr_1rr, showing a duplication subtending Detarioideae, (b) cluster1032_1rr_1rr, showing a duplication subtending Papilionoideae, (c) cluster1248_1rr_1rr and (d) cluster2941_1rr_1rr, both with a duplication subtending the legume family. Trees for (e) cluster51_7rr_1rr and (f) cluster544_1rr_1rr show evidence of more than one duplication, including one specific to Papilionoideae in the former.

Figure S2. Numbers of gene duplications mapped across the species tree as estimated by Phyparts. The topology used is the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,038 homolog clusters.

Numbers above branches (with blue background) and below branches (with yellow background) represent numbers of duplications and numbers of homolog trees with duplications without or with a bootstrap filter of 50%, respectively.

Figure S3. Numbers of gene duplications mapped across the species tree as estimated by Notung. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,324 homolog clusters.

Figure S4. Numbers of gene duplications mapped across a non-binary species tree as estimated by Notung. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes, with poorly supported relationships collapsed. duplications were counted from 8,324 homolog clusters.

Figure S5. Root-to-tip lengths per taxon with partitions of fixed local clocks indicated. Pruned taxa with outlier root-to-tip lengths are indicated with an X, partitions are indicated with colors. (a) FLC3, (b) FLC6, (c) FLC8.

Figure S6. Chronogram estimated under the UCLN clock model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S7. Chronogram estimated under the UCLN clock model, with alternative prior 2. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S8. Chronogram estimated under the RLC model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S9. Chronogram estimated under the FLC3 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S10. Chronogram estimated under the FLC6 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S11. Chronogram estimated under the FLC8 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S12. Chronogram estimated under the FLC8 model, with alternative prior 1. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles, with alternative calibrations as red circles.

Figure S13. Chronogram estimated under the STRC model. Numbers behind nodes indicate 95% HPD intervals. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S14. Substitution rates as estimated in FLC8 analyses for the different clock partitions. Boxplots for each partition for (a) alternative prior 1 and (b) the “normal” prior setting. Colors correspond to the partitions as shown in Figures 5, S14, S15 and S18.

Figure S15. Histograms of age estimates of duplication nodes, for (a) the duplications mapped to the legume crown node in the Notung analysis and for duplication nodes in gene trees with only (b) Detarioideae, (c) Caesalpinioideae and (d) Papilionoideae included.

Chapter III

HYBRID CAPTURE OF 964 NUCLEAR GENES GENERATES A ROBUST BACKBONE PHYLOGENY FOR THE MIMOSOID CLADE (LEGUMINOSAE, CAESALPINIOIDEAE), YET FAILS TO RESOLVE THE HYPERFAST, SPECIES-RICH, PANTROPICAL INGIOID RADIATION

Authors:
Erik J.M. Koenen[1], Catherine Kidner[2,3], James Nicholls[2], R. Toby Pennington[3,4], Luciano P. de Queiroz[5], Melissa Luckow[6], Gwylim Lewis[7] and Colin E. Hughes[1]

[1] Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstrasse 107, CH-8008, Zurich, Switzerland
[2] School of Biological Sciences, University of Edinburgh, King's Buildings, Mayfield Rd, Edinburgh, UK
[3] Royal Botanic Gardens Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, U.K.
[4] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX4 4RJ, U.K.
[5] Departamento Ciências Biológicas, Universidade Estadual de Feira de Santana, Avenida Transnordestina, s/n - Novo Horizonte, 44036-900, Feira de Santana, Brazil
[6] L.H. Bailey Hortorium, Department of Plant Biology, Cornell University, Ithaca, New York 14853 USA.
[7] Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, U.K.

Abstract

High-throughput sequencing of target enrichment DNA libraries is increasingly used to enhance phylogenetic resolution. Sequencing large numbers of nuclear gene markers also opens up opportunities to quantify gene tree discordance and explore the underlying reasons why some parts of the Tree of Life are difficult (or potentially impossible) to resolve. However, it is not clear how to best select the markers to target, nor how to assemble the resulting raw read data into orthologous loci. Here we explore these questions to generate a well resolved generic backbone phylogeny for the mimosoid legumes (Leguminosae: Caesalpinioideae-mimosoid clade), a prominent pantropical angiosperm clade of c. 3300 spp. of mostly trees and shrubs which occur abundantly across all major lowland tropical biomes. Previous mimosoid phylogenies have shown that none of the traditionally recognized mimosoid tribes are monophyletic, and revealed the non-monophyly of the largest genus, *Acacia*. Previous phylogenies also showed complete lack of resolution across the Ingioid clade, a large pantropical group comprising c. 38 genera and more than 2,000 species, within which generic delimitation remains in a state of considerable flux. We present a complete phylogenomics protocol using hybrid capture, from generating four newly sequenced transcriptomes spanning the mimosoid phylogeny, and selecting 964 targeted genes using a custom bioinformatics pipeline, to assembling and analysing the captured DNA sequence data. We enriched for and sequenced these genes for 115 mimosoids and 7 outgroup taxa using a custom target capture kit. We use maximum likelihood and Bayesian phylogenetic inference on concatenated alignments and a multi-species coalescent method to estimate the species tree. We also evaluate gene tree support and conflict and use phylogenetic network analysis to investigate phylogenetic signal across loci. We greatly improve phylogenetic resolution across the mimosoid phylogeny and show that the Ingioid clade can be resolved into four robustly supported larger and three smaller sub-clades, providing a framework for future reclassification. However, the deeper internodes along the backbone of this clade cannot be resolved with high certainty and nearly all loci show a lack of phylogenetic signal across that part of the phylogeny. We suggest this is mainly caused by hyperfast speciation during the early evolution of this clade, leading to ILS and poorly estimated gene trees lacking resolution. Further sampling is necessary to represent non-monophyletic and a few missing

genera, and to further enhance the mimosoid phylogeny, generate a robust new generic system for the Ingioid clade, and explore the timing and rates of diversification across the large pantropical Ingioid radiation.

Keywords: Hybrid capture, Leguminosae, Fabaceae, mimosoids, phylogenomics, hard polytomy

The field of plant phylogenetics has had tremendous impacts on botanical studies and taxonomic classification, macroevolution and biogeography, ever since pioneering studies inferred the first molecular phylogenies of plant taxa (Chase et al., 1993). While those early studies used only a single locus, the plastid gene *rbcL*, modern studies often employ hundreds to several thousands of genes to infer phylogenetic relationships (e.g. Lee et al., 2011; Wen et al., 2013; Wickett et al., 2014; Yang et al., 2015; Zeng et al., 2017). Many recent plant phylogenetic studies use some form of targeted enrichment (Cronn et al., 2012; Lemmon & Lemmon, 2013), and hybrid capture is now one of the most widely used methods for phylogenomics (e.g. Mandel et al., 2014; Weitemier et al., 2014; Nicholls et al., 2015; Jones & Good, 2016; Sass et al., 2016; Johnson et al. 2018; Couvreur et al., 2019; Ojeda et al., 2019). Methods for selecting genes (e.g. Vatanparast et al., 2018; Johnson et al., 2018) and assembling and analysing the captured DNA sequence data are rapidly emerging. For example, a number of pipelines are available to assemble gene matrices from the captured loci (Yang & Smith, 2014; Johnson et al., 2016; Moore et al. (2017), but each of these approaches has pros and cons, and there is no universally accepted protocol.

At the same time, it has become clear that many polytomies, i.e. parts of the Tree of Life that are difficult to resolve, are rife with conflicting gene tree histories, caused by lack of phylogenetic signal (Salichos & Rokas, 2013; Shen et al., 2017), incomplete lineage sorting (ILS), hybridization and/or horizontal gene transfer, or combinations of these (Rokas et al., 2003; Salichos & Rokas, 2013; Suh et al., 2015; Copetti et al., 2017; Moore et al., 2017; Walker et al., 2018; Koenen et al., to be resubmitted). Many of these issues are particularly associated with rapid episodes of species divergence and diversification, i.e. evolutionary radiations. Detailed analyses of phylogenetic signal and conflict across large numbers of gene

trees can shed light on what factors are causing lack of resolution and determine whether they should be represented, in extreme cases, as candidate hard polytomies (Suh , 2016), i.e. episodes of (nearly) instantaneous speciation of three or more lineages.

Here, we use hybrid capture to enrich a set of 964 putative low copy genes with the goal of inferring a robust generic backbone phylogeny for the mimosoid legumes, which include a large clade which has been particularly recalcitrant to phylogenetic resolution. The mimosoid clade (LPWG, 2017), formerly subfamily Mimosoideae, comprises c. 3,300 species in c. 87 genera of trees, shrubs, geoxyles and lianas. Highly characteristic of the clade is the diversity of inflorescence types composed of many small flowers where the colourful stamens are the most conspicuous floral whorl, the inflorescence is the unit of pollinator attraction, and in many genera pollen is aggregated into, often large, polyads. In turn, flower morphology is relatively uniform across mimosoids, with mainly quantitative variation in sizes of organs, numbers of floral parts and the degree of fusion within whorls. Based on a few conspicuous floral characters, the clade has been divided into three large tribes: Mimoseae (10 or fewer free stamens), Acacieae (usually > 30 free stamens) and Ingeae (usually > 30 stamens partly fused into a tube), which have all been shown to be non-monophyletic (Fig.1) (Luckow et al., 2003, Luckow, 2005; LPWG, 2013). The smaller tribe Parkieae is also non-monophyletic and *Parkia* itself is nested within Mimoseae (Luckow et al., 2003), as is the monotypic tribe Mimozygantheae (Luckow et al., 2005). With a dysfunctional tribal classification, generic affinities have increasingly been referred to informally-named clades (e.g. Hughes et al., 2003), and informal generic groups (Lewis et al., 2005) or alliances (Barneby & Grimes, 1996). Many genera also remain poorly defined and have been shown, or are suspected, to be non-monophyletic, and generic delimitation remains in a state of considerable flux, frustrated by what appears to be rampant morphological homoplasy and lack of phylogenetic resolution. This has been especially the case for tribe Ingeae, where different authors have proposed starkly discordant generic systems (see Table 1 in Brown, 2008). In particular, the genus *Albizia* is poorly defined and its delimitation remains one of the most challenging taxonomic problems in the legume family.

Most species of mimosoids occur in the tropics, with major centres of diversity in Central and South America, Australia, Africa and Madagascar. Mimosoids occur in virtually

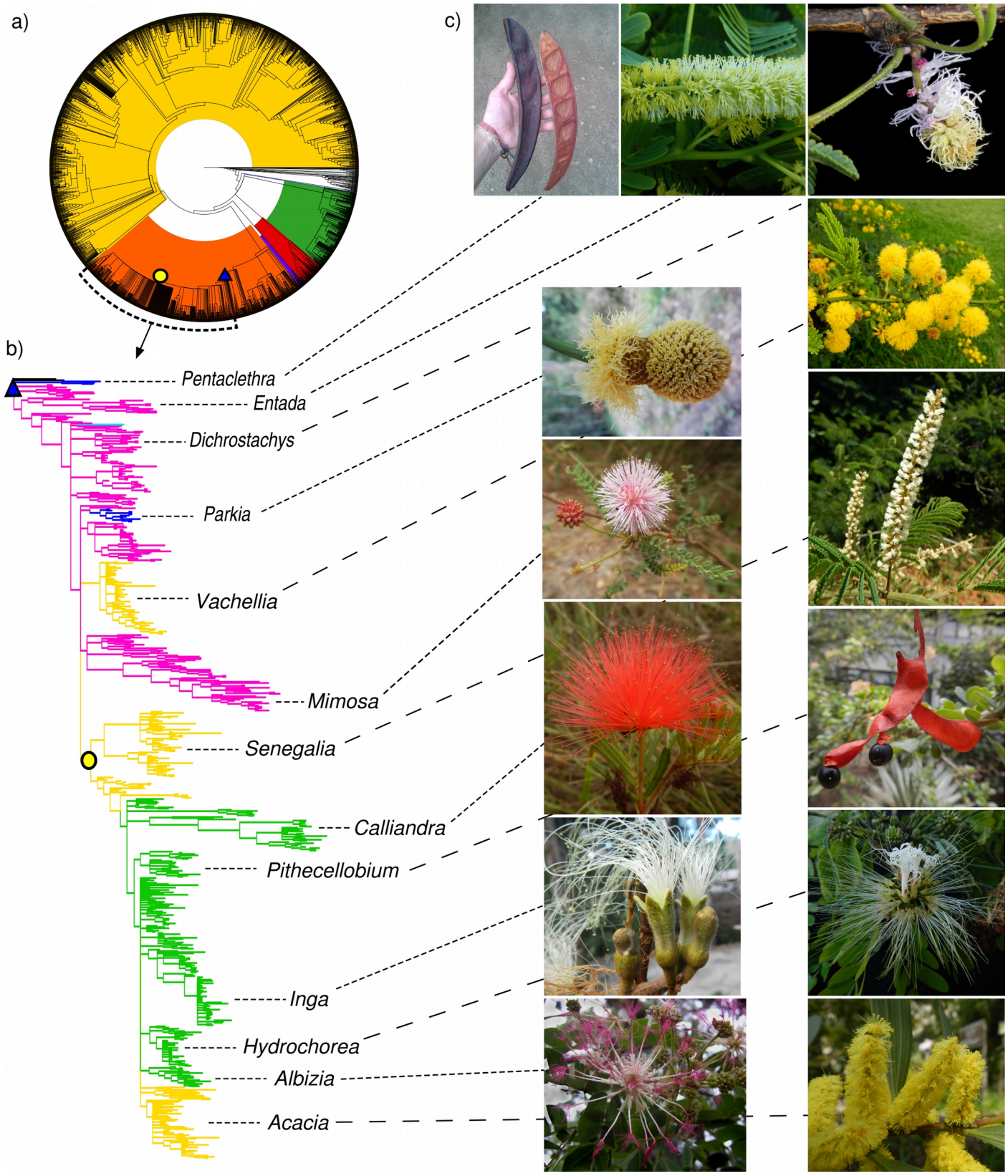

a)
b)
c)
Pentaclethra
Entada
Dichrostachys
Parkia
Vachellia
Mimosa
Senegalia
Calliandra
Pithecellobium
Inga
Hydrochorea
Albizia
Acacia

Figure 1. Mimosoid phylogeny, classification and diversity. a) Majority-rule bootstrap consensus tree from 1000 bootstrap replicates of the matK phylogeny from LPWG (2017), indicating the position of the mimosoid clade (crown node indicated by a blue triangle) within subfamily Caesalpinioideae (shaded orange) and showing that the Ingioid clade (crown node indicated by a yellow circle) is the least resolved portion of the legume phylogeny. b) Majority-rule Bayesian consensus tree for the mimosoid clade, extracted from the matK phylogeny of LPWG (2017), highlighting the non-monophyly of mimosoid tribes Parkieae (dark blue), Mimoseae (pink), Acacieae (yellow) and Ingeae (green). The monotypic Mimozygantheae (light blue) is nested in Mimoseae. c) From left to right, top to bottom: Pod valves of *Pentaclehtra macrophylla*, spicate inflorescence of *Entada chrysostachys*, heteromorphic inflorescences of *Dichrostachys akataensis* and likewise for *Parkia bahiae*, compound inflorescence of *Vachellia karroo*, capitate inflorescence of *Mimosa blanchetii*, spicate inflorescence of *Senegalia ataxacantha*, capitate inflorescence of *Calliandra fuscipila*, dehisced fruit with seeds suspended on arillodia in *Pithecellobium diversifolium*, flowers of *Inga subnuda*, dimorphic inflorescence with enlarged central flowers of *Hydrochorea corymbosa* and likewise for *Albizia grandibracteata*, spicate inflorescences of *Acacia longifolia*. All photos by EK.

every lowland tropical biome or vegetation type (except mangroves) and are abundant and diverse in both seasonally dry habitats (savannas and seasonally dry forests of the succulent biome sensu Ringelberg et al., submitted) where they often dominate, and wet forests, where certain clades are ubiquitous, e.g. the genus *Inga* and allies in American tropical wet forests where the genus *Inga* is thought to represent a recent rapid species radiation (Richardson et al., 2001). Mimosoids have apparently re-invented themselves numerous times to inhabit either rainforests, savannas or seasonally dry tropical forests (SDTFs sensu Pennington 2000 & 2009; the succulent biome sensu Gagnon et al. 2019; Ringelberg et al. submitted). Because of their prominence in diverse tropical lowland biomes, the mimosoid clade offers an excellent study system to investigate adaptation along the gradient from ever-wet to seasonally dry and arid tropical climates, as well as the extent of phylogenetic biome conservatism vs biome shifting. However, a well-resolved species tree for comparative analyses is lacking. Lack of resolution is particularly stark in the large Ingeae and Acacieae p.p. clade (Figs. 1a & b) (Luckow et al., 2003; Miller et al., 2003; Brown et al., 2008; Bouchenak-Khelladi et al., 2010; LPWG, 2017), hereafter referred to as the Ingioid clade. This clade includes some 2,000 species in c. 38 genera, but the relationships among these genera are largely unknown, even although all were sampled in the most comprehensive legume

phylogeny to date (Fig. 1) (LPWG, 2017). In fact, this clade appears to represent the least resolved part of the legume phylogeny as a whole (Fig. 1a).

In this study, we present a complete phylogenomics project using hybrid capture, from generating transcriptome data and selecting targeted genes to assembling and analysinganalysing the captured DNA sequence data for a large set of accessions of the mimosoid clade. The targeted genes were selected using a custom pipeline, which has recently also been used to select loci for other taxonomic groups (Couvreur et al., 2019; Ojeda et al., 2019) and which is potentially useful across all taxonomic groups. Using these genome-scale data we generate a robust generic backbone phylogeny for the mimosoid clade. We focus especially on the large, poorly resolved Ingioid clade, to try to understand why relationships in this clade have been so contentious. To address this question, we also investigate (the strength of) conflicting signals across gene trees and use phylogenetic network approaches to assess whether the evolution of the Ingioid clade is tree-like or polytomous.

Methods

This study describes a complete phylogenomics project using hybrid capture, from generating transcriptome data and selecting targeted genes (Fig. 2) to assembling and analysinganalysing the captured DNA sequence data (Fig. 3) for a large set of accessions of the mimosoid clade.

RNAseq to generate genomic resources

With no fully sequenced genome for mimosoids available when we started this study, we generated transcriptome data for four mimosoid genera to select nuclear markers for targeted enrichment. For the species *Albizia julibrissin, Entada abyssinica* and *Microlobius foetidus*, seedlings were grown at the Botanic Garden of the University of Zurich, and RNA extracted from young leaves and shoot tips, as well as roots (*A. julibrissin*), using the RNeasy Mini kit (Qiagen). Libraries for sequencing were produced using the TruSeq RNA Library Prep

kit (Illumina) and sequenced 3-plex on an Illumina HiSeq-2000 sequencer, at the Functional Genomics Center in Zurich. Raw data were cleaned with prinseq-lite (Schmieder & Edwards, 2011), and transcriptomes assembled using Trinity (Grabherr et al., 2011; Haas et al., 2013) with default settings. In addition, transcriptome data for three species of *Inga* were generated at the Royal Botanic Gardens in Edinburgh following the same procedures. The separate assemblies for the three *Inga* species were combined into a non-redundant set of transcripts by running BLAST searches of the largest assembly against the second largest transcriptome assembly and adding all transcripts without a significant hit (e-value cut-off 1e-10) in the latter. This procedure was then repeated for the third species.

Selecting putative single copy genes

From the four transcriptome data sets, putative single- or low-copy nuclear genes were selected, using a procedure inspired by Wu et al. (2006) (Fig. 2). This procedure has recently also been used by Couvreur et al. (2019) and Ojeda et al. (2019), but as it was first designed for the mimosoid bait set, it is described in more detail here. First, for each of the four transcriptome data sets, TransDecoder was used to predict open reading frames (ORFs) and translate those to protein sequences, using default settings. Highly similar proteins were removed to reduce redundancy (that is, keeping only one protein sequence per gene and removing multiple alleles and isoforms) with CD-HIT (Li & Godzik, 2006). This was repeated with four cut-off values (90, 95, 97 and 99% identity), to avoid either clustering paralogous sequences with relatively low divergence, or keeping alleles and isoforms with relatively high divergence. This means that the following steps were each repeated four times, and from each repetition, only the putative orthologs that were more divergent among and less divergent within taxa were kept. For each transcriptome, we performed a BLAST search of the CD-HIT output against itself (“selfBLAST”) with an e-value cut-off of 1e-10, and sequences with multiple hits within the same transcriptome were removed to eliminate gene families. Next, a reciprocal best hit (RBH) algorithm was implemented in a custom python script, to compare the four transcriptome data sets after removing redundancy and gene families. This is an extension of the RBH triangulation method of Wu et al. (2006), where a set

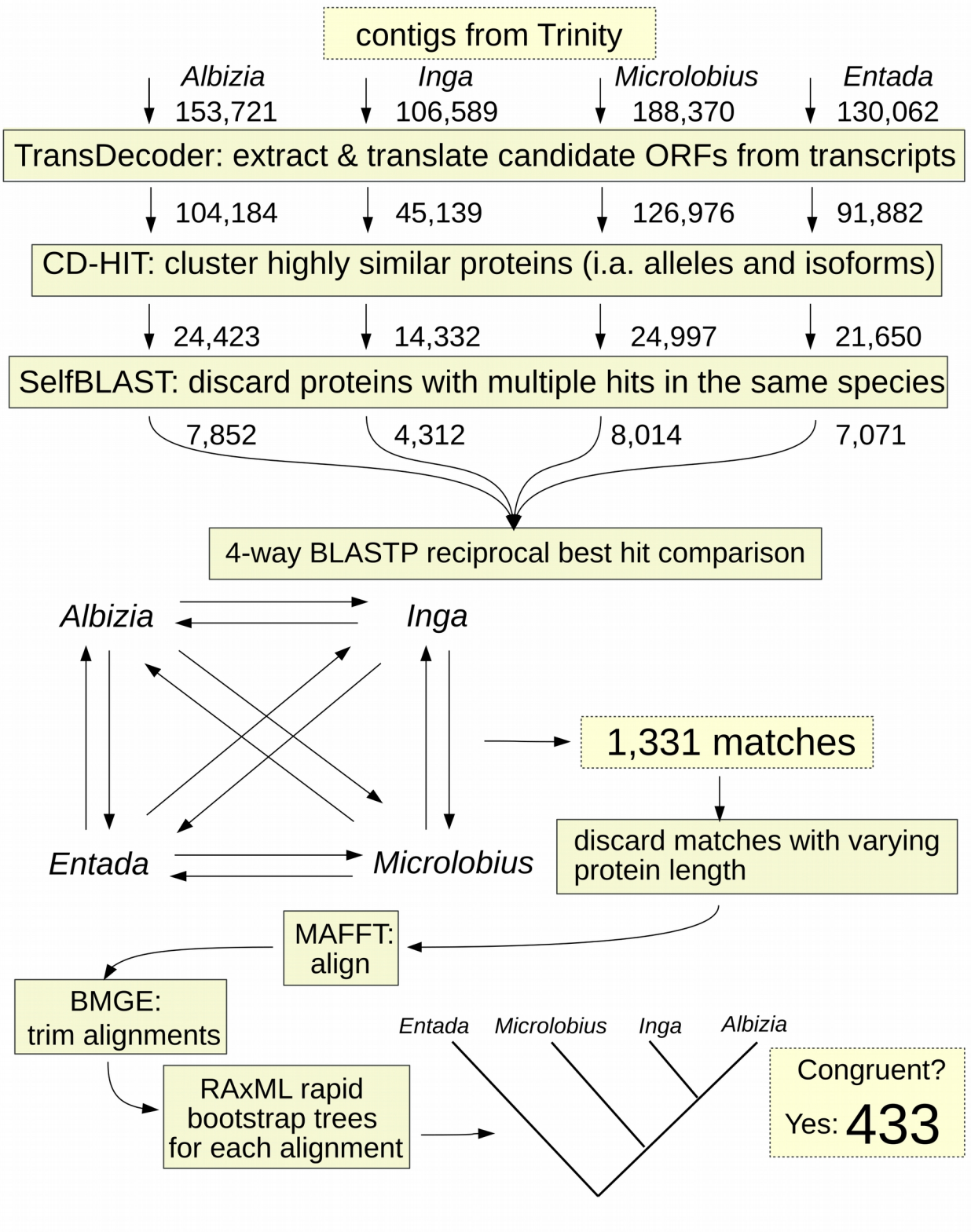


Figure 2. Target gene selection workflow, indicating the number of sequences and loci retained at each step.

of four sequences are considered as a putative ortholog if all possible pairwise reciprocal BLAST searches among the four transcriptomes yield the same RBH (Fig. 2). This works as follows: first, we take the first sequence of the transcriptome that we want to design the baits from (in our case *Albizia*) and run a BLAST search against one of the other transcriptomes; the best hit from the second transcriptome is then used as a query for a BLAST search against the first transcriptome, and when the original sequence that we started with is recovered as the best hit, this is considered an RBH. This is repeated for all combinations of transcriptomes by taking the sequence of the previous RBH and running a BLAST search against another transcriptome. This procedure was repeated for all sequences from the first transcriptome and then sequences from the four transcriptomes that gave an RBH across all pairwise BLAST searches which were written to separate FASTA files for each putative ortholog. Putative orthologs in which sequence length varied by more than 5% were discarded as an additional quality control step. From the resulting FASTA files, we performed a phylogenetic congruence test similar to that of Wu et al. (2006). Orthologs were aligned with MAFFT using the G-INS-i algorithm (Katoh et al., 2005), alignments trimmed with BMGE with default settings (Criscuolo et al., 2010) and rapid bootstrap analyses carried out with RAxML under the PROTCATLGF model (Stamatakis, 2014). If the resulting 95% bootstrap consensus topology was incongruent with the previously known, and well established relationships among the four taxa (Fig. 2), the putative ortholog was discarded. After running these procedures for each of the four different CD-HIT cut-off values, the resulting ortholog sets were combined as the ‘RBH4’ set.

Additionally, an ‘RBH3’ set was generated by comparing just the three largest transcriptomes (*Albizia*, *Entada* and *Microlobius*), but omitting the phylogenetic congruence test (because a minimum of four taxa are needed to infer a phylogeny). A third set of putative orthologs was generated by running RBH comparisons among the two largest transcriptomes (*Albizia* and *Microlobius*), to sets of genes found by De Smet et al. (2013) to be strictly or mostly single copy across 20 angiosperm genomes (using sequences of *Arabidopsis thaliana*). This third set is split into two subsets referred to as ‘SSC’ (strictly single copy) and ‘MSC’ (mostly single copy).

Bait design

For bait design, the sequences of the *Albizia julibrissin* transcriptome were used, as the genus *Albizia* and allied genera are the focus of an ongoing project in Zurich, and this will increase successful capture for these taxa. We test the effectiveness of these baits across the mimosoid clade plus closely related genera of Caesalpinioideae. Intron-exon boundaries were predicted for all transcripts in the four ortholog sets (RBH4, RBH3, SSC and MSC), by running BLAST searches against a custom database combining the *Arabidopsis thaliana* (Lamesh et al., 2011), *Medicago truncatula* (Young et al., 2011) and *Glycine max* (Schmutz et al., 2010) genomes. For the genome database, gene models including introns were used, and the coordinates to which our transcripts aligned were used to partition sequences for each predicted exon to avoid designing baits spanning intron-exon boundaries. This step is not essential but is likely to increase the efficiency of the capture. In addition to coding sequences, we also included 120bp of the 3'-UTR and 240 bp of the 5'-UTR, but sequences obtained for these regions are not analysed further here. Furthermore, additional target genes were added that included functionally interesting genes and genes targeted for separate studies in *Inga* (Nicholls et al., 2015), but again, none of these genes are analysed here as we focus on the low copy loci selected for phylogenetic analysis. Final bait design was carried out by Mycroarray (now Arbor Biosciences), with 3x tiling, and RNA baits were synthesized as part of the myBaits Custom Target Capture kit.

DNA extraction, library preparation, hybrid capture and sequencing

We extracted DNA for 122 accessions, representing 74 of the c. 87 currently recognized mimosoid genera and 6 closely related genera of non-mimosoid Caesalpinioideae (voucher details in Table S1), using the Qiagen DNeasy Plant Mini Kit. Sequencing libraries were prepared with the NEBNext Ultra DNA Library Prep kit for Illumina (New England Biolabs), in combination with the NEBNext Multiplex Oligos for Illumina (both single and dual index kits). Libraries were quantified using qPCR and pooled prior to hybrid capture. Pools

consisted of 8-21 libraries based on approximate evolutionary distances to the species from which the baits were designed (e.g. species of *Albizia* were pooled together, species of closely related genera were pooled together in another pool, and species from more distantly related genera were pooled together in yet another pool, etc.). The different pools were then enriched for the targeted regions in separate reactions with the myBaits Custom Target Capture kit. Enriched pools were quantified and pooled into a single library that was sequenced on Illumina HiSeq 2000 at the Functional Genomics Center in Zurich.

Assembly of sequence data and aligned matrices

After demultiplexing, raw reads were processed with Trimmomatic (Bolger et al., 2014) to remove adapter sequence artefacts and trim or remove low quality reads (using the settings MAXINFO:40:0.1 LEADING:20 TRAILING:20), and PEAR (Zhang et al., 2013) to merge overlapping read pairs (after removing adapter artefacts but before trimming). Resulting fastq files of quality filtered merged, paired and unpaired reads were used in a *de novo* assembly for each accession using the SPAdes assembler (Bankevich et al., 2012). From the resulting scaffolds, we extracted all ORFs of at least 300bp long between two stop codons with getorf (using the option -find 2) from the Emboss software suite. We reduced redundancy in the set of ORFs found for each accession with cd-hit, using an identity cut-off of 0.99. For all ORFs from each accession a BLAST search was carried out against the target sequences and for each target a multifasta file was created. Each ORF for each accession was added to the target multifasta file for which it received the best BLAST hit under an e-value cut-off of 1e-10, resulting in multifasta files for each target with potentially multiple sequences per accession included, which we refer to hereafter as ‘clusters’ .

Numbers of reads on target were estimated by mapping the untrimmed reads to the bait sequences with BLAT (Kent, 2002), using a minimum sequence identity threshold of 70%. Numbers of recovered loci were estimated with BLASTX, using protein sequences for the 964 targeted genes as the database and the SPAdes contigs as the query sequences, with a e-value cutoff of 1e-10.

Using the fasta_to_tree.py script of Yang & Smith (2014), each cluster was aligned

with MAFFT, sites with excessive missing data were removed (with a minimum column occupancy of 0.3) and a tree was inferred for each cluster with RAxML. We used other scripts of Yang & Smith (2014) to trim outlier long tips (with relative and absolute cut-offs of 0.1 and 0.3, respectively), mask monophyletic and paraphyletic clusters belonging to the same taxon and cut deep paralogs (cutting internal branches above 0.3 and keeping subtrees of at least 25 accessions). From the resulting trimmed subtrees, new multifasta files were created for a second round of tree inference, trimming and masking. However, for the second round we used MACSE (Ranwez et al., 2018), instead of MAFFT, to obtain more accurate alignments and TreeShrink (with quantile of trees to remove set to q=0.1; Mai & Mirabab, 2018) to trim tips, instead of relative and absolute cutoffs. Finally, after cutting deep paralogs again, we extracted all non-overlapping sub-clusters with at least 25 accessions using the maximum inclusion (MI) method of Yang & Smith (2014).

One limitation of the ortholog selection pipeline modified from Yang and Smith (2014) and used here to assemble gene matrices from the captured loci, is that while it is well-suited for distinguishing orthologs from paralogs, it does not deal with fragmented sequences representing different exons in a very satisfying way. Other available pipelines can potentially deal with this issue, but have other limitations. For example, the Hybpiper pipeline (Johnson et al., 2016) could potentially reconstruct longer gene sequences than the method applied here, but does not automatically sort different paralogs into separate gene alignments. Similarly, Moore et al. (2017) described a method to classify exons by their respective paralog gene copies, which offers a promising approach, but it relies on having initial backbone gene family trees for all loci. However, since recombination may also take place in between different exons of the same gene, it is also possible that exons may be better evolutionary units to analyze than full gene sequences (Scornavacca & Galtier, 2017).

Besides analysing the targeted genes, we also extracted off-target reads with a BLAST hit against a chloroplast reference genome set of *Inga leiocalycina* (Dugas et al., 2015; Genbank accession KT428296), *Leucaena trichandra* (Dugas et al., 2015; Genbank accession KT428297) and *Erythrophleum fordii* (Huang et al., 2018; Genbank accession MG644609), assembled chloroplast sequences for all accessions, and extracted the coding sequences gene by gene using a custom python script with BLAST searches. The *clpP* gene

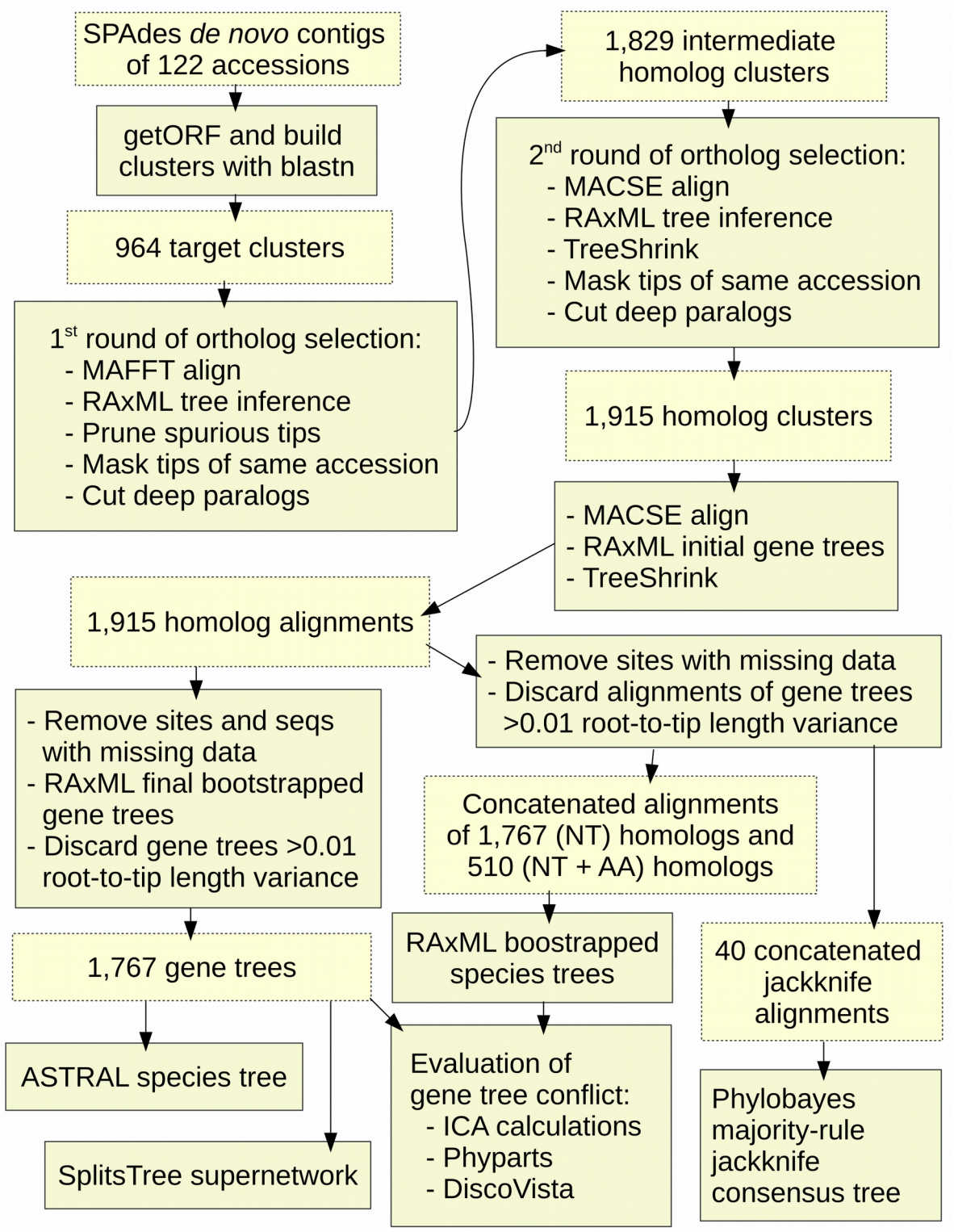


Figure 3. Workflow for phylogenetic ortholog selection and gene tree and species tree analyses.

was discarded because it shows accelerated evolution (Williams et al., 2015; Dugas et al., 2015), and yields a nonsensical gene tree. The *accD* gene has been lost from the chloroplast genome in several papilionoids, is highly variable in others (Magee et al., 2010), and is difficult to align across mimosoids so, we also removed this gene for phylogenetic analysis. The remaining 72 plastid genes were aligned with MACSE and concatenated with the pxcat program of the Phyx package.

Phylogenetics

The MI sub-clusters were aligned with MACSE (Ranwez et al., 2018) to yield codon alignments, codons with more than 95% missing data were removed using pxclsq from the Phyx package (Brown & Smith, 2017), and initial gene trees inferred with RAxML. Using TreeShrink with a relatively high quantile cut-off (q=0.25), we removed outlier long tips, to ensure a low error rate in the alignments. The drawback of this is that outgroup taxa and other taxa outside the "core mimosoids" (here defined by mrca of *Prosopis laevigata* and Ingeae/Acacieae – see below) also get pruned relatively frequently from these loci, but, given that the mimosoid phylogeny in those parts is already well-characterized from previous work (Luckow et al., 2003; Bouchenak-Khelladi et al., 2010), this is unlikely to be problematic.

For gene tree inference, codons with more than 75% missing data were removed from the alignments, after which sequences shorter than 300bp and at the same time occupying less than 50% of the total aligned length were removed. Gene trees were inferred with RAxML under the GTRGAMMA model with 200 rapid bootstrap replicates. Using pxlstr from the Phyx package root-to-tip variance was estimated to discover outlier gene trees that might have originated from poor orthology inference or alignment artefacts. After inspecting a subset of gene trees, we decided to discard all those with a root-to-tip variance > 0.01. Gene trees were used to calculate Internode Certainty All (ICA) values using RAxML (Kobert et al., 2016), for species tree analysis using ASTRAL-III (Zhang et al., 2018), and for phylogenetic supernetwork analysis. ASTRAL-III analyses were done on the best Maximum Likelihood (ML) gene trees, and subsets of gene trees with more than 25 or 50% of the accessions present to check if the analyses are sensitive to including gene trees with a lot of missing

data. We also ran the polytomy test in ASTRAL-III (Sayyari & Mirabab, 2018) to see for which nodes a polytomy null model could not be rejected.

Another way to analyze conflicting signals across gene trees is to infer a filtered Z-closure supernetwork (Whitfield et al., 2008). For deciding which splits to take into account we used the 'mintrees' parameter, allowing us to infer multiple networks including rarer splits or only fewer, more commonly observed, and therefore better supported, splits. For phylogenetic supernetwork analysis, we pruned all gene trees to a selection of taxa from the Ingioid clade representing all main lineages within it that were present in large proportions of the gene trees, yielding a total of 878 gene trees in which at least more than half of the selected Ingioid taxa were represented (at least 6 out of 11). All pruned gene trees with less than half of the selected taxa present were discarded. Phylogenetic supernetworks were constructed using Splitstree 4 (Huson, 1998), using different cut-offs for the MinTrees setting, representing 2.5, 5, 7.5 and 10% of the total number of gene trees.

For the concatenated alignments, codons that only had unambiguous characters for less than 10% of the total number of accessions were removed. Both nucleotide and translated peptide alignments of loci with more than half of the taxa present were concatenated with pxcat of the Phyx package. Loci for which the gene tree had a root-to-tip variance >0.01 were discarded prior to concatenation. Concatenated alignments, including the chloroplast alignment, were analysed with RAxML, using the GTRCAT model for DNA sequences and the PROTGAMMALG4X model for protein sequences (Le et al., 2008), running 200 rapid bootstrap replicates for each. Finally, we carried out a gene jackknifing analysis with Phylobayes (Lartillot et al., 2013) using the CATGTR model, by dividing the loci randomly over four relatively equally sized concatenated protein sequence alignments with 10 replicates, running a total of 40 analyses for 1000 cycles. For faster convergence, the ML estimate of the concatenated analysis in RAxML was provided as a starting tree for the chains. The first 500 cycles of each replicate were discarded as burn-in prior to summarizing a majority rule consensus tree over all replicates.

Visualizing gene tree discordance

Numbers of supporting and conflicting bipartitions for each node were extracted from the gene trees that had more than half the accessions present using Phyparts (Smith et al., 2015). For this, gene trees first had to be rooted, which was done using pxrr from the Phyx package, with a list of outgroup taxa outside the “core mimosoids” ranked by their relative divergence from Ingeae/Acacieae. Additionally, we visualize proportions of supporting and rejecting gene trees for selected clades with DiscoVista (Sayyari et al., 2018), from the same set of gene trees for which at least half the accessions are present.

Results

Transcriptome sequencing, gene selection and bait design

Transcriptome sequencing statistics are in Table 1; FASTQ files with raw read data are available on the European Nucleotide Archive (ENA), under accession numbers XXXX; assembled transcripts are available through the Transcriptome Shotgun Assembly (TSA) database, accession numbers XXXX. Results from the gene selection procedure are summarized in Fig. 2. After running the pipeline with four different similarity cut-offs in CD-HIT, we found 433 RBH4 and 334 RBH3 target genes. We recovered 320 MSC and 19 SSC genes, of which 134 and 8 genes respectively were already included in the RBH sets. Combining all gene sets we obtained a total of 964 low copy nuclear genes for enrichment. The complete coding sequences from the *Albizia julibrissin* transcriptome for these targeted genes are in Supplementary Data S1. The bait design included 24,856 probes at 3x tiling.

Targeted sequencing and data assembly

Sequencing and *de novo* assembly statistics for targeted sequencing for all accessions arepresented in Table S1. Accessions were enriched and sequenced in three separate batches, with different levels of multiplexing, which explains some of the variation

Table 1. Transcriptome sequencing and assembly

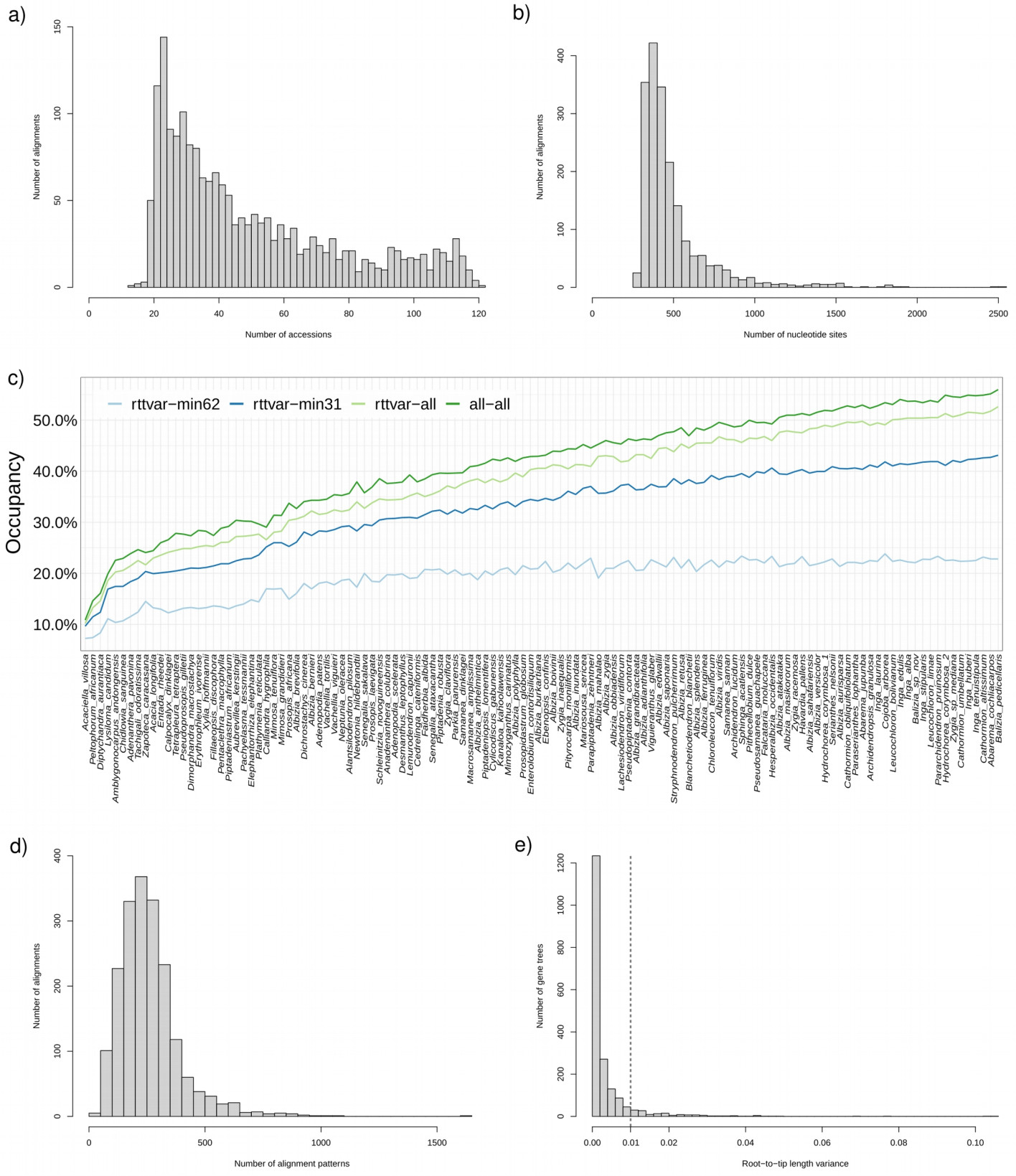
Taxon	Total number of reads	Quality filtered reads	Trinity contigs	Predicted ORFs
Albizia julibrissin	65,129,217	Left: 60,128,377 Right: 57,345,882	153,721	104,184
Entada abyssinica	65,006,875	Left: 59,821,838 Right: 56,882,422	130,062	91,882
Microlobius foetidus	97,515,912	Left: 89,669,576 Right: 85,024,338	188,370	126,976
Inga (3 species, non-redundant set)	NA	NA	106,589	45,139

observed in numbers of total reads and reads on target. Total read numbers varied from 1,360,502 to 70,271,424. For the largest batch of samples, the enrichment was less efficient with number of reads on target between 3.81 and 17.77%, while for the two smaller batches it varies between 69.00 and 85.27%. The percentage of reads on target is particularly low for taxa most distantly related to *Albizia julibissin* on which the bait sequences are based. Highly divergent sequences are not expected to be captured, but even so, these percentages may be exaggerated if the targeted sequences are highly divergent (<70% sequence identity to the baits), given the mapping threshold that we employed. Despite the variable enrichment efficiency, we were able to reconstruct at least partial sequences for the large majority of loci across almost all taxa (Table S1), with the number of target loci recovered, as determined by BLASTX searches of the scaffolds, ranging from 644 to 957.

After ortholog detection, a total of 1,915 gene alignments were recovered (Fig. 3), representing 767 of the targeted genes. This means that clusters representing the 197 remaining targeted genes were discarded because orthologous sub-clusters contained too few accessions, which may in turn be caused by poor phylogenetic resolution. For 279

targets, only a single gene alignment was recovered, i.e. they are putatively single copy. For the remainder of the gene alignments it is sometimes difficult to establish whether the multiple alignments represent paralogous copies, multiple exon alignments for the same gene that became separated during phylogenetic ortholog detection, or gene alignments that were split into two taxon sets because of long internal branches. Using BLAST searches of the longest sequence of each gene alignment against the target sequences, it became clear that many of these do indeed represent different fragments (most likely exons) of the same gene. Furthermore, some of the multiple alignments for the same gene do not have any overlapping accessions, suggesting they represent orthologous sequences for two distinct groups of taxa. It is thus not straightforward to accurately determine the precise number of paralog copies among the targeted genes.

The number of accessions per gene alignment ranged from 13 to 121, although there are not many alignments with fewer than 25 taxa, since this was the threshold used for extracting gene alignments from the homolog clusters, prior to running TreeShrink with a more stringent quantile percentage cutoff and removing sequences with a high proportion of missing data (Fig. 4a). At the same time, all gene trees estimated here are partial with respect to the species tree, and the numbers of accessions per alignment are skewed towards fewer than half the accessions being present. Aligned length per gene alignment varied from 282 to 2,526 bp, with few alignments below the threshold used to extract the alignments (300 bp) and a skew towards relatively short alignments (Fig. 4b). Taxon occupancy per locus shows a c. 4-fold difference, with generally higher occupancy for members of the Ingioid clade compared to more divergent taxa (Fig. 4c). However, even the least represented accession (*Acaciella villosa*), is still present in 274 gene alignments, which is likely sufficient to resolve its placement in the phylogeny, at least in concatenated analyses. Numbers of distinct alignment patterns, an indication of the phylogenetic informativeness of an alignment, show an uneven distribution across gene alignments, suggesting there are potentially only a few highly informative genes in the data set, but also few that are relatively uninformative. However, this does not indicate whether certain genes are particularly informative for deeper nodes or for more recent ones.

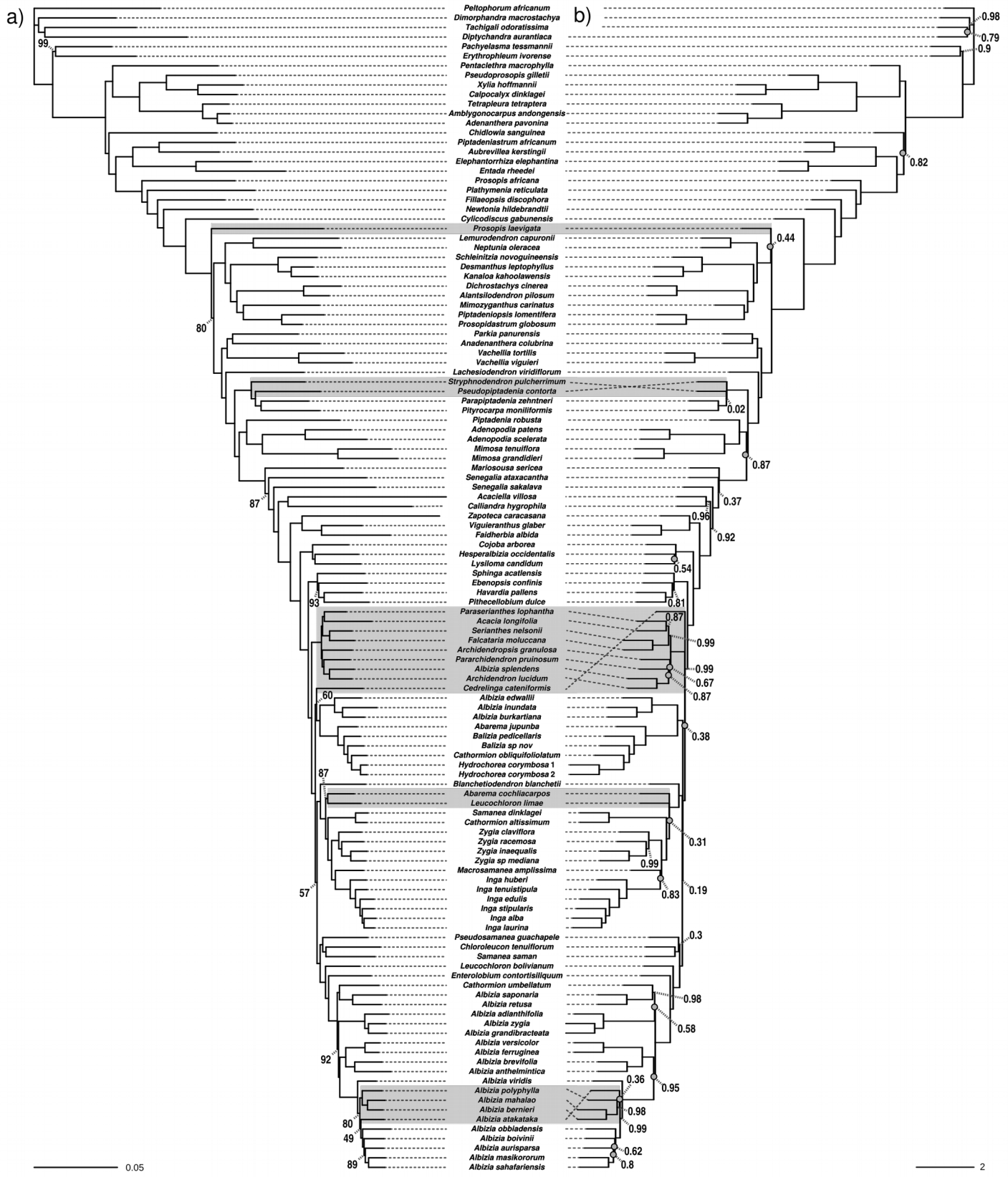

a)
Number of alignments
Number of accessions
b)
Number of alignments
Number of nucleotide sites
c)
rttvar−min62
rttvar−min31
rttvar−all
all−all
Occupancy
50.0%
40.0%
30.0%
20.0%
10.0%
d)
Number of alignments
Number of alignment patterns
e)
Number of gene trees
Root−to−tip length variance

(previous page) Figure 4. Statistics for recovered loci. a) Number of accessions per locus, b) aligned length per locus, c) taxon occupancy per locus (all vs rttvar = all loci or only those with <0.01 root-to-tip variance in the gene trees; all vs min31 vs min62 = without or with minimum taxon cut-offs of 25 or 50%, respectively), d) number of alignment patterns per locus and e) root-to-tip variance in the inferred gene trees, with the dashed line at 0.01 indicating the cut-off for retaining or discarding loci.

Gene and species tree inference

Gene trees inferred from the gene alignments showed that 148 genes had relatively high root-to-tip variances (>0.01; Fig. 4e). Inspection of these gene trees suggests they are not suitable for phylogenetic reconstruction due to very marked branch length variation, making the inferred relationships unreliable. These gene trees were discarded and not analysed further. After excluding these loci, the remaining 1,767 loci were aligned giving a total aligned length of 861,525 bp, with 450,375 alignment patterns and 62.12% missing data. A second concatenated alignment for only those loci with at least half of the accessions included (510 genes/exons), has a total aligned length of 254,250 bp, or 84,750 amino acids with 176,713 or 73,179 alignment patterns, respectively, and 34.89% missing data. Jackknife alignments consist of between 127 and 129 genes with total aligned lengths of 19,949 to 22,218 amino acids. The chloroplast alignment is 60,321 bp long, contains 16,589 alignment patterns and has only 17.33% missing data.

The concatenated ML and ASTRAL species tree analyses yielded highly supported and highly similar topologies, except for a relatively small number of internodes (Figs 5 & 6). ML analyses of the concatenated alignment of 510 loci (Figs. S1 & S2) show higher support and almost identical topologies. The Bayesian jackknife consensus tree (Fig. 6) shows a polytomy at the base of the mimosoid clade, involving the position of *Chidlowia* and several polytomies within the Ingioid clade, including a large one along the backbone of that clade. The chloroplast phylogeny (Fig. S3), confirms that sequence data for the chloroplast genome can be efficiently extracted and analysed from off-target reads in hybrid capture experiments (as shown by Weitemier et al., 2014). The plastid topology differs in some places from the species trees inferred from nuclear gene data, but is also less robustly supported, particularly


a)
b)
Peltophorum africanum
Dimorphandra macrostachya
Tachigali odoratissima
Diptychandra aurantiaca
Pachyelasma tessmannii
Erythrophleum ivorense
Pentaclethra macrophylla
Pseudoprosopis gilletii
Xylia hoffmannii
Calpocalyx dinklagei
Tetrapleura tetraptera
Amblygonocarpus andongensis
Adenanthera pavonina
Chidlowia sanguinea
Piptadeniastrum africanum
Aubrevillea kerstingii
Elephantorrhiza elephantina
Entada rheedei
Prosopis africana
Plathymenia reticulata
Fillaeopsis discophora
Newtonia hildebrandtii
Cylicodiscus gabunensis
Prosopis laevigata
Lemurodendron capuronii
Neptunia oleracea
Schleinitzia novoguineensis
Desmanthus leptophyllus
Kanaloa kahoolawensis
Dichrostachys cinerea
Alantsilodendron pilosum
Mimozyganthus carinatus
Piptadeniopsis lomentifera
Prosopidastrum globosum
Parkia panurensis
Anadenanthera colubrina
Vachellia tortilis
Vachellia viguieri
Lachesiodendron viridiflorum
Stryphnodendron pulcherrimum
Pseudopiptadenia contorta
Parapiptadenia zehntneri
Pityrocarpa moniliformis
Piptadenia robusta
Adenopodia patens
Adenopodia scelerata
Mimosa tenuiflora
Mimosa grandidieri
Mariosousa sericea
Senegalia ataxacantha
Senegalia sakalava
Acaciella villosa
Calliandra hygrophila
Zapoteca caracasana
Vigueranthus glaber
Faidherbia albida
Cojoba arborea
Hesperalbizia occidentalis
Lysiloma candidum
Sphinga acatlensis
Ebenopsis confinis
Havardia pallens
Pithecellobium dulce
Paraserianthes lophantha
Acacia longifolia
Serianthes nelsonii
Falcataria moluccana
Archidendropsis granulosa
Pararchidendron pruinosum
Albizia splendens
Archidendron lucidum
Cedrelinga cateniformis
Albizia edwallii
Albizia inundata
Albizia burkartiana
Abarema jupunba
Balizia pedicellaris
Balizia sp nov
Cathormion obliquifoliolatum
Hydrochorea corymbosa 1
Hydrochorea corymbosa 2
Blanchetiodendron blanchetii
Abarema cochliacarpos
Leucochloron limae
Samanea dinklagei
Cathormion altissimum
Zygia claviflora
Zygia racemosa
Zygia inaequalis
Zygia sp mediana
Macrosamanea amplissima
Inga huberi
Inga tenuistipula
Inga edulis
Inga stipularis
Inga alba
Inga laurina
Pseudosamanea guachapele
Chloroleucon tenuiflorum
Samanea saman
Leucochloron bolivianum
Enterolobium contortisiliquum
Cathormion umbellatum
Albizia saponaria
Albizia retusa
Albizia adianthifolia
Albizia zygia
Albizia grandibracteata
Albizia versicolor
Albizia ferruginea
Albizia brevifolia
Albizia anthelmintica
Albizia viridis
Albizia polyphylla
Albizia mahalao
Albizia bernieri
Albizia atakataka
Albizia obbiadensis
Albizia boivinii
Albizia aurisparsa
Albizia masikororum
Albizia sahafariensis
0.05
2

(previous page) Figure 5. Generic backbone phylogeny of mimosoid legumes. Comparison between the concatenated ML and ASTRAL species trees, with grey shading indicating topological differences. a) RAxML tree inferred from the full concatenated alignment (1,767 loci) with bootstrap support indicated for internodes that received less than 100%, and branch lengths in number of substitutions per site. b) ASTRAL species tree inferred from 1,229 loci with more than a quarter of the accessions present, with branch lengths in coalescent units. Local posterior probability is indicated for internodes that received less than 1, circles on nodes indicate those nodes for which a polytomy could not be rejected. Terminal branch lengths in the ASTRAL tree are set at 1 (instead of 0) for better visualization.

within the Ingioid clade. We do not discuss the chloroplast phylogeny further, but note it's potential utility to study maternal inheritance in hybrids. All alignments and trees are included in Supplementary Data files S1-6, and TreeBase, accession number XXXXX.

Characterization of well-supported clades

The species trees provide evidence for 14 well-defined clades (Fig. 6) that receive high support in (almost) all analyses, and most of them are also well-supported across gene trees (Fig 7a).

The **Xylia clade** is defined as the clade that includes all genera from the monophyletic Adenanthera group (Lewis et al., 2005) plus *Pentaclethra.* The clade includes two distinctive sub-clades: 1) the sub-clade of *Xylia, Pseudoprosopis* and *Calpocalyx,* which is restricted to Africa and Madagascar and characterized by sickle-shaped explosively dehiscent fruits; and 2) the sub-clade comprising *Adenanthera*, *Tetrapleura* and *Amblygonocarpus*, with the former genus largely restricted to South-East Asia, and characterized by indehiscent or non-explosively dehiscent fruits. Since *Pentaclethra* (sister to these two sub-clades, see Fig. 6) also has explosively dehiscent fruit, this is likely the ancestral fruit state of the Xylia clade.

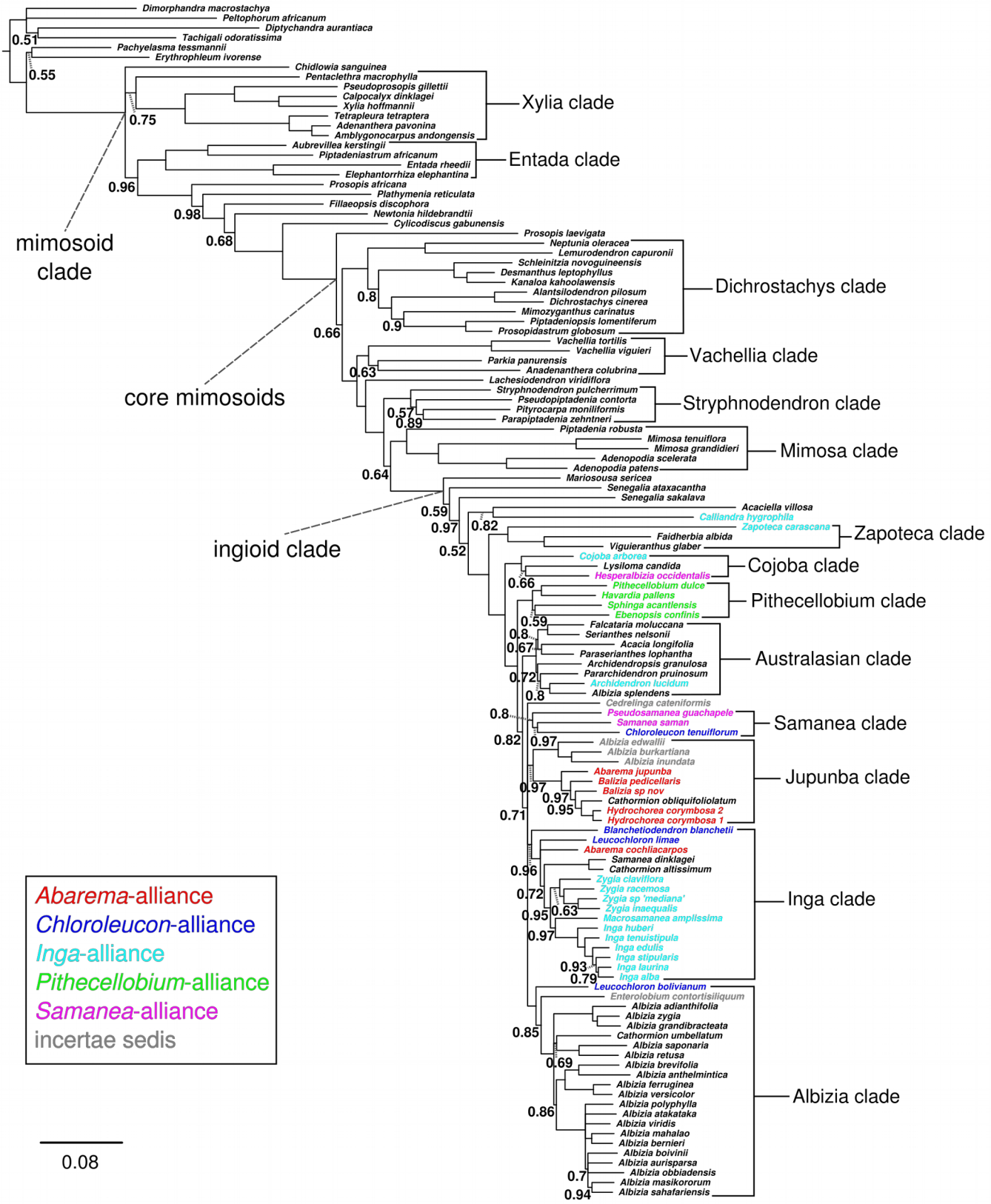
The **Entada clade** includes the genera *Entada, Elephantorrhiza, Piptadeniastrum* and *Aubrevillea*, and has its centre of diversity in, and three of the four genera restricted to continental Africa. *Entada* is more widespread, with several species in Madagascar and a few species with drift-seeds having attained large pantropical distributions following trans-oceanic dispersal.

Core mimosoids are here defined as the clade that includes the mrca of *Prosopis laevigata, Dichrostachys cinerea* and *Inga edulis*. This clade includes the bulk of mimosoid species and all of the larger genera. While there are no obvious morphological synapomorphies for the clade, it is subtended by a particularly long internode (Fig. 6), hence leading us to recognize the clade here.

The **Dichrostachys clade** includes the informal Dichrostachys and Leucaena groups (Hughes et al., 2003; Lewis et al., 2005), as well as *Mimozyganthus*, *Piptadeniopsis* and *Prosopidastrum* (Luckow et al., 2005), and the genera *Neptunia* and *Lemurodendron.* Most taxa in this clade are found in the succulent biome (Ringelberg et al., submitted) comprising seasonally dry tropical forest and thorn scrub (SDTF sensu Pennington et al., 2000 & 2009), with centres of diversity in Mexico and Central America (the Leucaena group) and Madagascar (the Dichrostachys group). *Lemurodendron* is monotypic, endemic to Madagascar, known until recently from only a handful of herbarium collections from the 1960s, and of previously unknown phylogenetic affinities. We re-collected it in NW Madagascar in 2014 allowing us to include it here. The sister-group relationship of *Lemurodendron* with *Neptunia* is perhaps surprising given their disparate morphologies, but arguably *Neptunia* is morphologically unlike any other mimosoid due to its (semi-) aquatic lifestyle. While not restricted to it, nor universal within it, the presence of heteromorphic inflorescences with showy staminodes at the base, is highly characteristic of this clade.

The **Vachellia clade** consists of the genera *Vachellia*, *Anadenanthera* and *Parkia.* While these genera do not share any conspicuous morphological features, the clade is well supported in all analyses.

The informal Piptadenia group (sensu Lewis and Elias, 1981; and Luckow, , minus the genus *Anadenanthera*) is here resolved into two well supported clades, the **Stryphnodendron clade** and the **Mimosa clade**. The former includes *Parapiptadenia, Pityrocarpa, Pseudopiptadenia, Stryphnodendron,* and *Microlobius* (not sampled here) (Simon et al., 2015). The Mimosa clade includes *Adenopodia*, *Mimosa* and *Piptadenia.* The monotypic genus *Lachesiodendron,* recently segregated from *Piptadenia* (Ribeiro et al., 2018), is placed outside both these clades and instead forms the sister-group of the remainder of the Piptadenia group and the Ingioid clade.


Dimorphandra macrostachya
Peltophorum africanum
Diptychandra aurantiaca
Tachigali odoratissima
Pachyelasma tessmannii
Erythrophleum ivorense
Chidlowia sanguinea
Pentaclethra macrophylla
Pseudoprosopis gillettii
Calpocalyx dinklagei
Xylia hoffmannii
Tetrapleura tetraptera
Adenanthera pavonina
Amblygonocarpus andongensis
Xylia clade
Aubrevillea kerstingii
Piptadeniastrum africanum
Entada rheedii
Elephantorrhiza elephantina
Entada clade
Prosopis africana
Plathymenia reticulata
Fillaeopsis discophora
Newtonia hildebrandtii
Cylicodiscus gabunensis
mimosoid clade
Prosopis laevigata
Neptunia oleracea
Lemurodendron capuronii
Schleinitzia novoguineensis
Desmanthus leptophyllus
Kanaloa kahoolawensis
Alantsilodendron pilosum
Dichrostachys cinerea
Mimozyganthus carinatus
Piptadeniopsis lomentiferum
Prosopidastrum globosum
Dichrostachys clade
Vachellia tortilis
Vachellia viguieri
Parkia panurensis
Anadenanthera colubrina
Vachellia clade
Lachesiodendron viridiflora
Stryphnodendron pulcherrimum
Pseudopiptadenia contorta
Pityrocarpa moniliformis
Parapiptadenia zehntneri
Stryphnodendron clade
core mimosoids
Piptadenia robusta
Mimosa tenuiflora
Mimosa grandidieri
Adenopodia scelerata
Adenopodia patens
Mimosa clade
Mariosousa sericea
Senegalia ataxacantha
Senegalia sakalava
Acaciella villosa
Calliandra hygrophila
Zapoteca carascana
Faidherbia albida
Vigueranthus glaber
Zapoteca clade
ingioid clade
Cojoba arborea
Lysiloma candida
Hesperalbizia occidentalis
Cojoba clade
Pithecellobium dulce
Havardia pallens
Sphinga acantlensis
Ebenopsis confinis
Pithecellobium clade
Falcataria moluccana
Serianthes nelsonii
Acacia longifolia
Paraserianthes lophantha
Archidendropsis granulosa
Pararchidendron pruinosum
Archidendron lucidum
Albizia splendens
Australasian clade
Cedrelinga cateniformis
Pseudosamanea guachapele
Samanea saman
Chloroleucon tenuiflorum
Samanea clade
Albizia edwallii
Albizia burkartiana
Albizia inundata
Abarema jupunba
Balizia pedicellaris
Balizia sp nov
Cathormion obliquifoliolatum
Hydrochorea corymbosa 2
Hydrochorea corymbosa 1
Jupunba clade
Blanchetiodendron blanchetii
Leucochloron limae
Abarema cochliacarpos
Samanea dinklagei
Cathormion altissimum
Zygia claviflora
Zygia racemosa
Zygia sp 'mediana'
Zygia inaequalis
Macrosamanea amplissima
Inga huberi
Inga tenuistipula
Inga edulis
Inga stipularis
Inga laurina
Inga alba
Inga clade
Leucochloron bolivianum
Enterolobium contortisiliquum
Albizia adianthifolia
Albizia zygia
Albizia grandibracteata
Cathormion umbellatum
Albizia saponaria
Albizia retusa
Albizia brevifolia
Albizia anthelmintica
Albizia ferruginea
Albizia versicolor
Albizia polyphylla
Albizia atakataka
Albizia viridis
Albizia mahalao
Albizia bernieri
Albizia boivinii
Albizia aurisparsa
Albizia obbiadensis
Albizia masikororum
Albizia sahafariensis
Albizia clade
Abarema-alliance
Chloroleucon-alliance
Inga-alliance
Pithecellobium-alliance
Samanea-alliance
incertae sedis
0.08

(previous page) Figure 6. Well-defined and highly supported clades in the mimosoid phylogeny. Clades are annotated on the Bayesian jackknife majority-rule consensus tree, with posterior probability values for internodes with less than 1.00 pp indicated. Coloured taxon names indicate non-monophyly of all but one of the alliances recognized by Barneby & Grimes (1996), as per the legend.

The **Ingioid clade** is well-supported in all analyses and is defined as the clade that includes all genera of tribe Ingeae plus *Acacia* and all its segregates except *Vachellia* (Fig. 6). Like the core mimosoids, it is subtended by a relatively long internode. All taxa in the clade share the feature of flowers with more than 10 stamens, which is otherwise only present in *Vachellia*. Fusion of the stamens into a tube is exclusively found in this clade and characterizes most of the genera. Within this notoriously difficult and poorly-resolved clade, in this study, we are able to recognize several well-supported sub-clades for the first time.

The **Zapoteca clade** includes the genera *Faidherbia*, *Viguieranthus* and *Zapoteca,* as well as *Sanjappa* and *Thailentadopsis* (Souza et al., 2016), which are not sampled here. The clade has a pantropical distribution. Typical for this clade are the fruits elastically dehiscent from the apex, similar to those of *Calliandra*, and only lacking in *Faidherbia.* Spinescent stipules are found in *Faidherbia*, *Sanjappa* and *Thailentadopsis*.

The **Pithecellobium clade** is identical to the Pithecellobium alliance of Barneby & Grimes (1996). The clade is native to the Americas, has its center of diversity in Mexico and Central America, and is characterized by spinescent stipules.

The **Cojoba clade** apart from *Cojoba* itself, includes *Lysiloma* and *Hesperalbizia* and is also native to the Americas, centered in Central America and the Caribbean.

The **Australasian clade** includes the largest genus of mimosoids, the predominantly Australian *Acacia* s.s., the large genus *Archidendron* that is widespread in South-East continental Asia, Malesia and the Pacific, as well as *Archidendropsis, Falcataria, Pararchidendron, Paraserianthes* and *Serianthes.* The species *Albizia splendens*, previously segregated as *Serialbizzia* along with one other species of *Albizia* (…; Nielsen …), is nested in this clade*. Wallaceodendron* (not sampled here), is also tentatively included in this clade. Because the clade is almost entirely restricted to Australasia and the Pacific region, its informal name refers to its biogeography.

The **Jupunba clade** is largely composed of the Abarema alliance of Barneby &

Grimes (1996), with the exclusion of the type species of *Abarema* (*A. cochliacarpos* which is nested in the Inga clade), and hence we name the clade after *A. jupunba,* the epithet of which is also a later synonym of *Abarema* and the likely generic name for transferring most of the *Abarema* species (Iganci et al., 2016). The American species of the genus *Albizia,* placed in section *Arthrosamanea* by Barneby & Grimes (1996), are included here and form the sister group to rest of the clade. Apart from the exclusively Neotropical Abarema alliance and *Albizia* sect. *Arthrosamanea*, the African *Cathormion obliquifoliolatum* is also nested in this clade. Together with *C. rhombifolium* (not sampled, also African)*, C. obliquifoliolatum* is morphologically highly similar to the Neotropical *Hydrochorea*, with the water-dispersed seed presumably having facilitated trans-atlantic dispersal. Dimorphic flowers are found in the majority of species across the clade.

The **Inga clade** includes the Neotropical genera *Blanchetiodendron*, *Inga*, *Leucochloron* (except *L. bolivianum*), *Macrosamanea* and *Zygia,* and is particularly diverse in lowland rainforests. *Blanchetiodendron* is a highly distinctive genus and is sister to the rest of the clade. Two African taxa, *Cathormion altissimum* and *Samanea dinklagei* are also nested in this clade for which a new genus will be needed. While not sampled here *Enterolobium* sect. *Robrichia* is also included in this clade (Elvia Souza, personal communication). Two species of that section, *Blanchetiodendron* and *C. altissimum* are the only taxa in this clade with dimorphic flowers.

The **Samanea clade** includes *Chloroleucon*, *Pseudosamanea* and *Samanea*, is restricted to the Neotropics, and all genera have species with dimorphic flowers.

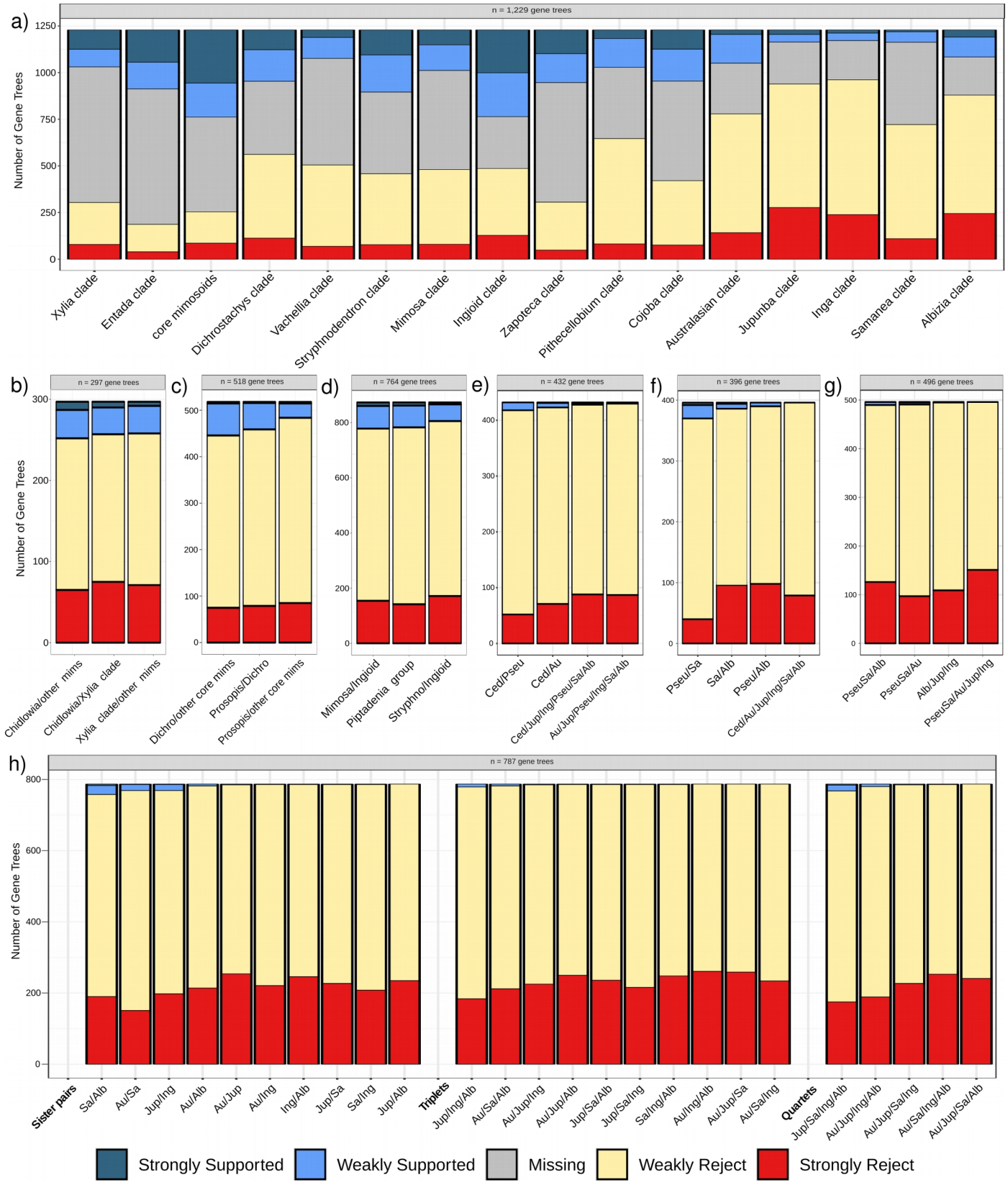
The **Albizia clade** includes the African, Malagasy and SE-Asian species currently accommodated in *Albizia* (except *A. splendens*)*,* plus the Neotropical *Enterolobium* pro parte (Elvia de Souza, unpubl. data) and *Leucochloron bolivianum* which is shown to be unrelated to the rest of *Leucochloron* (Fig. 6). Furthermore, the type species of *Cathormion*, *C. umbellatum* is nested within *Albizia* and should therefore be synonymized with that genus. Both *Albizia* and *Enterolobium* include species with dimorphic flowers.

Evaluation of support for inferred relationships

The ASTRAL topology differs in only five places from the ML topology (Fig. 5): 1) *Prosopis laevigata* is sister to the Dichrostachys clade with 0.44 pp, instead of to the rest of the core mimosoids with 80% BS, 2) *Stryphnodendron pulcherrimum* and *Pseudopiptadenia contorta* have swapped positions, with 0.02 pp in the ASTRAL tree, while the alternative relationship in the ML tree has full support, 3) *Cedrelinga cateniformis* is sister to a large clade composed of several subclades of the Ingioid clade with 0.38 pp, instead of being sister to the Jupunba clade with 60% BS, 4) *Abarema cochliacarpos* and *Leucochloron limae* are not sister taxa, with full support, while they are in the ML tree with 87% BS and 5) *Albizia atakataka* is in a different position in the two trees, with 49% BS versus 0.36 pp.

Support along the backbone of the phylogeny, with the exception of the Ingioid clade, is generally high in concatenated analyses (Figs. 5 & 6), but taking into account conflicting signals across gene trees, levels of support are less robust. Many internodes receive relatively low ICA support (<0.5; Fig. S4) suggesting significant conflict for those nodes. In a few cases, ICA values below zero indicate that the most common conflicting bipartitions are more prevalent than the supporting ones. Comparing proportions of gene tree bipartitions supporting an internode, relative to the most common conflicting bipartions, all other conflicting bipartitions, and uninformative gene trees, it is clear that the majority of gene trees are either uninformative or contain an infrequent conflicting bipartition (Fig. S5). This strongly suggests that the majority of gene trees lack phylogenetic signal, especially across the Ingioid backbone.

The ASTRAL polytomy test showed that for several nodes, the null model of a polytomy could not be rejected (p-value = 0.05; Figs 4b and S6). We evaluated three questionable deeper nodes along the backbone of the phylogeny (placement of *Chidlowia*, placement of *Prosopis* and mono/paraphyly of the Piptadenia group) as well as along the backbone of the Ingioid clade, where polytomies cannot be ruled out for several nodes. This shows that the placement of *Chidlowia* as sister to all other mimosoids excluding the Xylia clade is preferred slightly over the two alternative hypotheses (Fig. 7b). For *Prosopis*, a sister group relationship with the rest of the core mimosoids is equally or slightly better supported


a)
n = 1,229 gene trees
Number of Gene Trees
Xylia clade
Entada clade
core mimosoids
Dichrostachys clade
Vachellia clade
Stryphnodendron clade
Mimosa clade
Ingioid clade
Zapoteca clade
Pithecellobium clade
Cojoba clade
Australasian clade
Jupunba clade
Inga clade
Samanea clade
Albizia clade
b)
n = 297 gene trees
Chidlowia/other mims
Chidlowia/Xylia clade
Xylia clade/other mims
c)
n = 518 gene trees
Dichro/other core mims
Prosopis/Dichro
Prosopis/other core mims
d)
n = 764 gene trees
Mimosa/Ingioid
Piptadenia group
Stryphno/Ingioid
e)
n = 432 gene trees
Ced/Pseu
Ced/Au
Ced/Jup/Ing/Pseu/Sa/Alb
Au/Jup/Pseu/Ing/Sa/Alb
f)
n = 396 gene trees
Pseu/Sa
Sa/Alb
Pseu/Alb
Ced/Au/Jup/Ing/Sa/Alb
g)
n = 496 gene trees
PseuSa/Alb
PseuSa/Au
Alb/Jup/Ing
PseuSa/Au/Jup/Ing
h)
n = 787 gene trees
Number of Gene Trees
Sister pairs
Sa/Alb
Au/Sa
Jup/Ing
Au/Alb
Au/Jup
Au/Ing
Ing/Alb
Jup/Sa
Sa/Ing
Jup/Alb
Triplets
Jup/Ing/Alb
Au/Sa/Alb
Au/Jup/Ing
Au/Jup/Alb
Jup/Sa/Alb
Jup/Sa/Ing
Sa/Ing/Alb
Au/Ing/Alb
Au/Jup/Sa
Au/Sa/Ing
Quartets
Jup/Sa/Ing/Alb
Au/Jup/Ing/Alb
Au/Jup/Sa/Ing
Au/Sa/Ing/Alb
Au/Jup/Sa/Alb
Strongly Supported
Weakly Supported
Missing
Weakly Reject
Strongly Reject

(previous page) Figure 7. Evaluation of gene tree support for selected nodes. a) Bargraphs of supporting and rejecting gene trees for the clades identified in this study and for alternative topologies involving b) the placement of *Chidlowia*, c) placement of *Prosopis*, d) monophyly or paraphyly of the Piptadenia group, e) placement of *Cedrelina*, f) placement of *Pseudosamanea*, g) affinities of the Samanea clade and h) all possible sister pairs, clade triplets and quartets within the polytomous portion of the Ingioid clade after pruning *Cedrelinga* and the Samanea clade from the gene trees. Note that for panels b)-h), the bars for each graph are sorted from most to least supported. Abbreviations: mims = mimosoids, Stryphno = Stryphnodendron clade, Ced = *Cedrelinga cateniformis*, Pseu = *Pseudosamanea guachapele*, Au = Australasian clade, Jup = Jupunba clade, Ing = Inga clade, Sa = *Samanea saman* and *Chloroleucon tenuiflorum*, Alb = Albizia clade, PseuSa = Samanea clade.

across gene trees than the two alternatives (Fig. 7c). For the Piptadenia group, paraphyly is slightly more often supported across gene trees than monophyly, with the Mimosa clade as the most likely sister group of the Ingioid clade (Fig. 7d). Within the Ingioid clade, there is a notable lack of resolution especially in the clade that includes *Cedrelinga* plus the Australasian, Jupunba, Inga, Samanea and Albizia clades. The phylogenetic placement of the monotypic *Cedrelinga* appears to be unstable with hardly any gene tree support for any of its possible placements (Fig. 7e). There are some weakly supporting gene trees showing a sister-group relationship of *Cedrelinga* with *Pseudosamanea*, but that taxon is more likely related to *Chloroleucon* and *Samanea* (Fig. 7f), and one of the other three possible placements (Fig. 7e) is probably more likely. There are no gene trees strongly supporting *Cedrelinga* as sister to the rest, and for the other two options there is just one gene tree strongly in support of each. A sister-group relationship between the Samanea and Albizia clades has minimal support across gene trees, even though it is found in several species tree analyses, and remains the most likely possibility relative to alternatives (Fig. 7g).

These results suggest that *Cedrelinga* and *Pseudosamanea*, and perhaps also the other two genera of the Samanea clade, are potentially causing lack of resolution in the Ingioid clade, acting as “rogue taxa”, for example due to lack of phylogenetic signal for the placement of these taxa or long branch attraction (LBA) artefacts, particularly for *Cedrelinga*. Another possibility is that ancient hybridization has occurred, giving rise to (some) of these rogue lineages. ML analyses on the concatenated alignment of 510 genes omitting these taxa does indeed increase support along the Ingioid backbone (compare Figs. S2, S7-S8). To

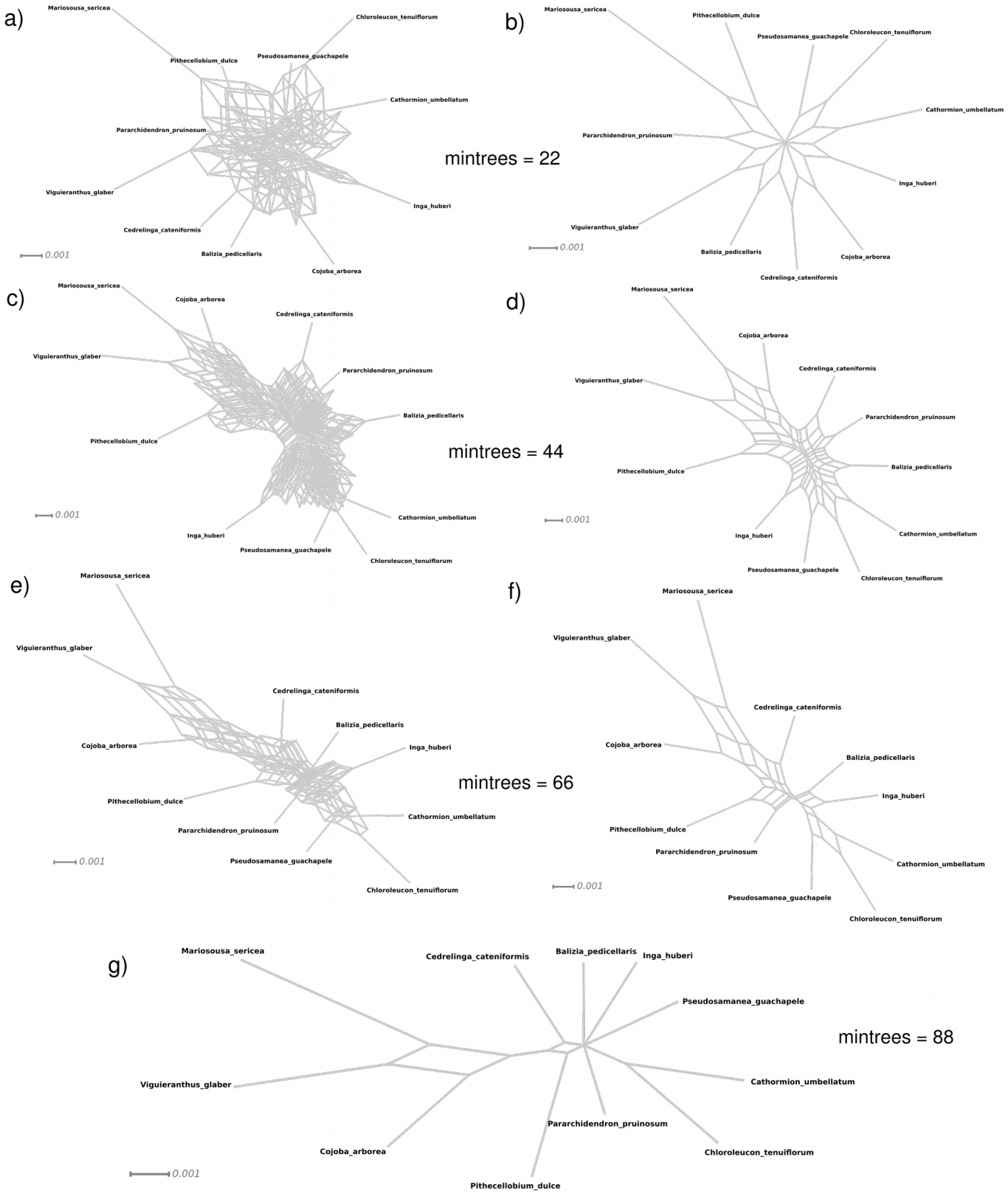
investigate this further, we evaluated support for all possible groupings of the Australasian, Jupunba, Inga, Samanea and Albizia clades as sister clades, triplets and quartets across gene trees with *Cedrelinga* and *Pseudosamanea* removed. This shows that the sister-group relationship of the Albizia and Samanea clades (with only *Cloroleucon* and *Samanea* included) is more likely than any other conflicting relationship (Fig. 7h) and that the Jupunba and Inga clades are likely to be sister clades. No well supported triplets are found, while the quartet that unites the Jupunba, Inga, Samanea and Albizia clades is better supported than all other possible quartets (Fig. 7h). Taken together, this would suggest a branching order of (Australasian((Jupunba,Inga),(Samanea,Albizia)) for these clades. However, none of the possible relationships among these clades, nor the placements of *Cedrelinga* and *Pseudosamanea*, appear in many gene trees with strong support, and it is striking that there are many more strongly conflicting gene trees for most of these.

Phylogenetic supernetwork analysis

At the lowest mintrees setting (n=22, c. 2.5% of the total number of trees; Figs. 8a & b), there appears to be little signal. Increasing to n=44 or 66 (Figs. 8c-f), the network becomes somewhat more treelike and shows more or less the same relationships among clade representatives as the gene tree support summarization (Figs. 7e-h). However, increasing mintrees to n=88 causes that resolution to collapse (Fig. 8g), showing how just limited phylogenetic signal hints at a resolved topology.

Discussion

In this study, we show that targeted enrichment, or more specifically hybrid capture, is a powerful and effective way to reconstruct the phylogeny of a challenging taxonomic group, in line with findings across a rapidly growing number of other groups s (e.g. Mandel et al., 2014; Weitemier et al., 2014; Nicholls et al., 2015; Jones & Good, 2016; Sass et al., 2016; Johnson et al. 2018; Couvreur et al., 2019; Ojeda et al., 2019). The phylogenetic resolution and statistical support obtained here offer a significant improvement over previous phylogenies (Luckow et al., 2003; Bouchenak-Khelladi et al., 2010; LPWG, 2017), yielding a


a)
Mariosousa_sericea
Chloroleucon_tenuiflorum
Pithecellobium_dulce
Pseudosamanea_guachapele
Cathormion_umbellatum
Pararchidendron_pruinosum
Viguieranthus_glaber
Inga_huberi
Cedrelinga_cateniformis
Balizia_pedicellaris
Cojoba_arborea
0.001
mintrees = 22
b)
Mariosousa_sericea
Pithecellobium_dulce
Pseudosamanea_guachapele
Chloroleucon_tenuiflorum
Cathormion_umbellatum
Pararchidendron_pruinosum
Inga_huberi
Viguieranthus_glaber
Balizia_pedicellaris
Cojoba_arborea
Cedrelinga_cateniformis
0.001
c)
Mariosousa_sericea
Cojoba_arborea
Cedrelinga_cateniformis
Viguieranthus_glaber
Pararchidendron_pruinosum
Balizia_pedicellaris
Pithecellobium_dulce
Cathormion_umbellatum
Inga_huberi
Pseudosamanea_guachapele
Chloroleucon_tenuiflorum
0.001
mintrees = 44
d)
Mariosousa_sericea
Cojoba_arborea
Cedrelinga_cateniformis
Viguieranthus_glaber
Pararchidendron_pruinosum
Pithecellobium_dulce
Balizia_pedicellaris
Inga_huberi
Cathormion_umbellatum
Pseudosamanea_guachapele
Chloroleucon_tenuiflorum
0.001
e)
Mariosousa_sericea
Viguieranthus_glaber
Cedrelinga_cateniformis
Balizia_pedicellaris
Cojoba_arborea
Inga_huberi
Pithecellobium_dulce
Cathormion_umbellatum
Pararchidendron_pruinosum
Pseudosamanea_guachapele
Chloroleucon_tenuiflorum
0.001
mintrees = 66
f)
Mariosousa_sericea
Viguieranthus_glaber
Cedrelinga_cateniformis
Cojoba_arborea
Balizia_pedicellaris
Inga_huberi
Pithecellobium_dulce
Pararchidendron_pruinosum
Cathormion_umbellatum
Pseudosamanea_guachapele
Chloroleucon_tenuiflorum
0.001
g)
Mariosousa_sericea
Cedrelinga_cateniformis
Balizia_pedicellaris
Inga_huberi
Pseudosamanea_guachapele
Viguieranthus_glaber
Cathormion_umbellatum
Pararchidendron_pruinosum
Cojoba_arborea
Chloroleucon_tenuiflorum
Pithecellobium_dulce
0.001
mintrees = 88

(previous page) Figure 8. Phylogenetic Z-closure filtered supernetworks with mintrees parameter set at 22, drawn a) with and b) without the Convex Hull algorithm, and the same for c) and d) mintrees setting at 44, e) and f) mintrees setting at 66, and finally, for the mintrees setting of 88 the networks with an without the Convex Hull method are identical, indicating that not many splits are included under this parameter setting.

robust generic backbone for the mimosoid clade (Figs. 5 & 6). Given the robust support across the phylogeny as a whole, the lack of resolution across the Ingioid backbone is striking. Furthermore, given the larg volume of DNA sequence data deployed in these analyses, this lack of resolution is probably not caused by insufficient data, but is instead most likely the result of extremely rapid speciation leading to a lack of phylogenetic signal and potentially some ILS, as implied by lack of resolution in a large majority of gene trees, and strongly supported conflict in a small fraction of the gene trees.

It is possible that fragmentation of exons from the same gene could have contributed to lack of resolution across gene trees and that other ortholog pipelines (see Methods) might yield longer alignments, avoid fragmentation of exons of the same gene, and hence be able to improve individual gene trees and thereby allow more accurate evaluation of alternative topologies. However, since recombination may also take place in between different exons of the same gene, suggesting that exons are better evolutionary units to analyze than full gene sequences (Scornavacca & Galtier, 2017). Either way, longer gene alignments are unlikely to yield much improved resolution in gene trees along the backbone of the Ingioid clade, since even the concatenated alignments do not find strong support in that part of the phylogeny. Therefore, there seems to be a strong lack of signal, ILS and/or ancient hybridization, leading to a potentially hard polytomy embedded within the Ingioid clade, involving 6 or 7 lineages.

The Ingioid clade is still recalcitrant to phylogenetic resolution

Within the Ingioid clade, relationships among well-supported subclades appear to be irresolvable (Fig. 7 e-h), which is surprising given the large number of genes deployed here, and raising the possibility that there may be a hard polytomy of six or seven lineages embedded within the Ingioid clade. While evaluation of supporting gene trees and the filtered supernetworks (Fig. 8) suggest some clade relationships as more likely than others, this may

simply be extracting the least conflicting signal from a data set where there is virtually no signal to begin with. In any case, there appears to be a large number of conflicting bipartitions among the set of gene trees (Fig. 7 e-h), and hardly any that strongly support any of the possible relationships among the Ingioid subclades.

Gene tree conflict is often attributed to ILS, as found in the initial radiation of the Neoaves clade of birds (Suh et al., 2015), which provides one of the most convincing examples of a hard polytomy documented so far (Suh, 2016), and which appears similarly irresolvable as the Ingioid clade. Suh et al. (2015) used retroposon insertion sites that are virtually free from homoplasy as strong evidence for ILS, while such evidence is lacking here. In other cases, such as mammals, ILS has been shown to be only a minor cause of gene tree conflict (Scornavacca & Galtier, 2017), suggesting that such conflict could equally be caused by gene tree estimation errors due to lack of phylogenetic signal, homoplasy, alignment errors and/or poor model fit (Richards et al., 2018). Across the Ingioid clade the majority of conflicting gene tree bipartitions appear to be rare and most of them are only weakly conflicting (Fig. 7 e-h). This suggests that most of the conflicting bipartitions stem from lack of phylogenetic signal, with gene tree estimation errors accounting in part for the strongly conflicting bipartitions (Richards et al., 2018). Other reasons for poor gene tree estimation include alignment errors, homoplasy, poor model fit and LBA artefacts. We have attempted to minimize alignment errors by using MACSE (Ranwez et al., 2018), which simultaneously aligns coding sequences and the amino acid translations, yielding far better alignments than MAFFT making the additional computational time very worthwhile. The inter-related issues of homoplasy, poor model fit and LBA artefacts are less easily tackled and could be the main sources of gene tree estimation errors in our data set. In that case this conflict would constitute phylogenetic noise rather than genuine conflicting signal, and such noise is present across much of the tree (Fig. S5). However, even although the number of conflicting bipartitions for many nodes across the tree far outnumber the most prevalent bipartition (Fig. S5), none of the alternatives is close to equally prevalent in parts of the species tree where resolution and support are consistently high. Within the Ingioid radiation, however, there is simply not enough signal to override this noise. It is also possible that fragmentation of exons from the same gene could have contributed to lack of resolution across gene trees and that

other ortholog pipelines (see Methods) might yield longer alignments, avoid fragmentation of exons of the same gene, and hence be able to improve individual gene trees and thereby allow more accurate evaluation of alternative topologies. However, longer gene alignments are unlikely to yield improved resolution in gene trees along the backbone of the Ingioid clade, since even the concatenated alignments do not find strong support in that part of the phylogeny.

Apart from gene tree estimation errors and ILS, ancient hybridization during the radiation of the Ingioid clade could offer an alternative explanation for the large numbers of strongly conflicting gene tree topologies. The strong conflicting gene tree support for the placement of *Pseudosamanea* in particular could be indicative of hybridization, although it is also possible that LBA artefacts could be causing an apparent sister-group relationship with *Cedrelinga* in some gene trees.

For the Neoaves clade of birds, lack of treelike structure in phylogenetic supernetworks was very similar to that found in networks generated from simulated random topologies, suggesting this clade is indeed best considered a hard polytomy (Suh, 2016). Together with the lack of gene tree support (Figs 7 e-h), our supernetworks (Fig. 8) suggest that the Ingioid radiation perhaps also constitutes a hard polytomy. With intermediate mintrees parameter settings (Fig. 8 c-f) the networks show some structure. However, given that this resolution collapses at the higher mintrees setting (Fig. 8 g), this is likely driven by a very small number of gene trees, while conflicting gene trees largely outnumber the few supporting ones (Fig. 7 e-h), in line with the idea that many contentious relationships are supported by just a handful of genes (Shen et al., 2017). Our network at the lowest mintrees setting is very similar to that of a simulated hard polytomy (cf Figs. 8 a & b; with Fig. 4E in Suh, 2016). We therefore conclude, pending enhanced taxon sampling and eventually completely sequenced genomes, that there is potentially a hard polytomy embedded in the Ingioid clade, involving six or seven lineages, which is resistant to resolution even using sequences from 100s of nuclear genes. With complete exomes and positional homology data it will be possible to investigate the sorting of different unlinked exons, genes or other genomic elements (e.g. retroposons) across lineages within the Ingioid clade in greater detail to shed light on the underlying tree-like structure of the phylogeny, or the lack thereof.

The Ingioid radiation

The pervasive lack of phylogenetic signal in a large majority of genes across the persistent Ingioid polytomy, implies that initial divergence of this clade involved an episode of hyperfast speciation, i.e. an evolutionary radiation, that gave rise to a large pantropical clade of more than 2000 species. What combination of extrinsic environmental opportunities and intrinsic evolutionary trait innovations triggered this large radiation remains to be investigated in detail, but some initial hypotheses for testing are worth considering.

First, the prevalence across the Ingioid clade, of flowers with numerous elongated stamens partially fused into a tube and aggregated into capitate, often dimorphic inflorescences with enlarged central nectar producing flowers, accompanied by pollen transfer in large polyads, is associated with efficient pollination by animals such as bats, hummingbirds or sphingoid moths. This would promote outcrossing and tend to maintain relatively high effective population sizes. Second, lomentaceous fruits within the Ingioid clade are often associated with hydrochory, and fleshy or otherwise nutritious, often indehiscent, fruits with zoochory, both of these promoting efficient long-distance dispersal, facilitating expansion and spread of populations, and potentially prompting divergence and speciation. This combination of highly efficient pollination and effective seed dispersal could in part explain rapid speciation together with large effective population sizes leading to increased ILS.

Fossil evidence (Crepet & Dilcher, 1977; Crepet & Taylor, 1986) and previous time-calibrated legume phylogenies (Lavin et al., 2005; Koenen et al., to be resubmitted) suggest that mimosoids originated in the Early Eocene, with divergence of the core mimosoids during the Late Eocene or Oligocene. Fossil polyads of Australian *Acacia* s.s. from c. 23 Ma approximately at the Oligocene-Miocene boundary (Miller et al., 2013) constitute the oldest fossil evidence of a Ingioid genus and suggest that the radiation of the Ingioid clade most likely occurred in the Oligocene, and may therefore be related to evolutionary biotic turnover in this epoch due to global climatic cooling. At that time, connections between South-America, Australia and Antarctica could explain the concentrations of diversity in the Neotropics (most

Ingioid subclades) and Australasia (one large subclade). The robust generic backbone phylogeny generated here provides the starting point for investigating the timing, biogeographical history and macro-evolutionary trajectory of the mimosoids and especially the Ingioid radiation, when expanded with more dense taxon sampling.

Further work is needed to investigate the causes and consequences of this large radiation, whether it is associated with past environmental change, whether the mrca of the polytomous clade had particular traits that facilitated its nearly instantaneous radiation into 6 or 7 lineages, and how the massive morphological and ecological disparification associated with this clade played out from this ancestral explosion of lineages.

Implications for the taxonomic classification of mimosoids

The absence of a bifurcating topology could explain the widespread morphological homoplasy apparent across the Ingioid clade and the consequent difficulties associated with delimiting genera, the discordant generic systems of different authors, and the non-monophyly of generic groupings (of e.g. Barneby & Grimes), which were entirely morphologically based. For example, lomentaceous fruits that break up into one-seeded articles occur in at least six different lineages scattered across the Albizia, Inga and Jupunba clades plus *Cedrelinga cateniformis.* Dimorphic capitate inflorescences with an enlarged central nectar-producing flower are similarly phylogenetically scattered across genera in the Albizia, Jupunba and Samanea clades and in *Blanchetiodendron blanchetii*. While reconstructing the evolution of e.g. pollination and seed dispersal syndromes across the Ingioid clade would undoubtedly be illuminating in this regard, it remains unclear to what extent this will be possible in the face of lack of phylogenetic resolution. In this study, we have advanced our understanding of the evolutionary relationships among Ingioid legumes, moving forward from a soft polytomy that included almost all of the c. 40 genera in the clade, to identify a potentially hard polytomy that involves six or seven robustly supported lineages. These lineages provide a robust framework for recognizing a set of informally named sub-clades (Fig. 6), replacing the previously defined informal groups and alliances, most of which are now shown to be non-monophyletic.

This framework provides the first step towards a much needed new tribal (Linnean) / clade-based (Phylocode) classification of mimosoids. Achieving this will require expanded taxon sampling of all potentially non-monophyletic and missing genera within mimosoids, as well as wider sampling of genera across subfamily Caesalpinioideae as a whole, something that is currently being undertaken using the same gene set employed here (Ringelberg & Koenen, unpubl. data). However, it is already clear that establishing a Linnean classification of tribes for mimosoids would require recognition of a large number of monogeneric tribes because of the strong imbalance across the generic backbone phylogeny (Figs 5 & 6). This could prompt recognition of mimosoids as a single tribe within Caesalpinioideae, and a Phylocode classification to formally name and describe clades within the mimosoids along the lines outlined here, when the initial set of clades highlighted here is better characterized with denser taxon sampling.

At generic level, it has been clear for some time that despite significant progress, substantive further generic re-delimitation is needed across mimsoids to account for the non-monophyly of several genera (Luckow et al., 2003; Igançi et al., 2016; Ferm, unpubl data; De Souza et al., unpubl data). Our results add to this tally of non-monophyletic mimosoid genera. For example, while the non-monophyly of *Albizia* has been long suspected, we demonstrate robust support for at least three separate evolutionary lineages currently ascribed to the genus: *Albizia* s.s., which includes species from Africa, Madagascar and Asia; *Albizia* sect. *Arthrosamanea* which includes the Neotropical species; and *Albizia splendens*, formerly segregated together with *A. acle* (not sampled here) as *Serialbizzia* (Kostermans, 1954). Our results further suggest that *Balizia* is not monophyletic with respect to African *Cathormion obliquifoliolatum* and Neotropical *Hydrochorea*, providing further evidence that the genera of the Abarema alliance of Barneby & Grimes are in need of re-delimitation (Igançi et al., 2016). The non-monophyly of *Senegalia* (beyond the recent segregation of *Parasenegalia* and *Psueodsenegalia*) that was found with 100% BS or 1.00 pp in all analyses of nuclear data (Figs 5 & 6) is unexpected given that Boatwright et al. (2015) showed Malagasy *Senegalia* species grouping with the rest of the genus based on three chloroplast regions. Notably, in our chloroplast phylogeny, the two species of *Senegalia* form a sister pair with 100% BS (Fig. S3). Further analyses sampling more widely across *Senegalia* is necessary to assess

whether this is a case of genuine conflict between the nuclear and chloroplast genomes or not.

Concluding remarks

The idea that not all gene trees are identical to the species tree, in other words that they can be incongruent, has been around for a while (Maddison, 1997). What phylogenomics has been showing over the last decade, is that gene tree conflict caused by gene tree estimation error, ILS and/or hybridization is in fact extremely common, including for deep nodes in the tree of life (Doronina et al., 2015; Marcet-Houben & Gabaldón, 2015; Suh et al., 2015; Meier et al., 2017; Koenen et al., to be resubmitted). This means that phylogenomics is not only concerned with finding the most likely species tree, but perhaps more importantly with exploring the complexity of phylogenetic relationships of independent genomic elements such as exons, complete genes or syntenic blocks and how those relate to the evolution of species diversity and traits.

Funding

This work was supported by the Swiss National Science Foundation (Grant 31003A_135522 to C.E.H.), the Claraz Schenkung Foundation, and the Department of Systematic & Evolutionary Botany, University of Zurich.

Acknowledgements

We thank René Stalder, Markus Meierhofer and Manfred Knabe for greenhouse assistance, the K, L, NYBG, P and WAG herbaria for provision of leaf samples, Elvia Souza, Petala Ribeiro, Marcelo Simon, Joao Iganci, Marli Morim and Francis Bonadeu for help with fieldwork in Brazil and the staff of the Kew Madagascar Conservation Center and The Botanical and Zoological Garden of Tsimbazaza for support during fieldwork in Madagascar, the Functional Genomics Center Zurich (FGCZ), especially Catherine Aquino, for lab support and high-throughput sequencing and the S3IT of the University of Zurich for the use of the ScienceCloud computational infrastructure.

Author contributions

EK and CEH designed the study. EK did the labwork, analyses and wrote the first version of the manuscript. CK helped with bioinformatics, LPQ, ML and GL contributed tissue samples for sequencing, CK, JN, RTP contributed data. EK, LPQ, ML and CEH did the fieldwork. All co-authors contributed to writing the final manuscript.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403-410.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. and Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5), pp.455-477.

Barneby, R.C. and Grimes, J., 1996. Silk tree, guanacaste, monkey's earring: a generic system of the synandrous Mimosaceae of the Americas. Part I. *Abarema*, *Albizia*, and allies. *Memoirs of the New York Botanical Garden*, 74(1).

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114-2120.

Bouchenak-Khelladi, Y., Maurin, O., Hurter, J. and Van der Bank, M., 2010. The evolutionary history and biogeography of Mimosoideae (Leguminosae): an emphasis on African acacias. *Molecular Phylogenetics and Evolution*, 57(2), pp.495-508.

Brown, G.K., 2008. Systematics of the tribe Ingeae (Leguminosae-Mimosoideae) over the past 25 years. *Muelleria*, 26(1), pp.27-42.

Brown, G.K., Murphy, D.J., Miller, J.T. and Ladiges, P.Y., 2008. *Acacia* ss and its relationship among tropical legumes, tribe Ingeae (Leguminosae: Mimosoideae). *Systematic Botany*, 33(4), pp.739-751.

Brown, J.W., Walker, J.F. and Smith, S.A., 2017. Phyx: phylogenetic tools for unix. *Bioinformatics*, 33(12), pp.1886-1888.

Cantino, P.D. and De Queiroz, K., 2000. PhyloCode: a phylogenetic code of biological

nomenclature.

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y.L. and Kron, K.A., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden*, pp.528-580.

Copetti, D., Búrquez, A., Bustamante, E., Charboneau, J.L., Childs, K.L., Eguiarte, L.E., Lee, S., Liu, T.L., McMahon, M.M., Whiteman, N.K. and Wing, R.A., 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proceedings of the National Academy of Sciences*, 114(45), pp.12003-12008.

Couvreur, T.L., Helmstetter, A.J., Koenen, E.J., Bethune, K., Brandão, R.D., Little, S.A., Sauquet, H. and Erkens, R.H., 2019. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. *Frontiers in Plant Science*, 9, p.1941.

Crepet, W.L. and Dilcher, D.L., 1977. Investigations of angiosperms from the Eocene of North America: a mimosoid inflorescence. *American Journal of Botany*, 64(6), pp.714-725.

Crepet, W.L. and Taylor, D.W., 1986. Primitive mimosoid flowers from the Paleocene-Eocene and their systematic and evolutionary implications. *American Journal of Botany*, 73(4), pp.548-563.

Criscuolo, A. and Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10(1), p.210.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V. and Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, 99(2), pp.291-311.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C., Maere, S. and Van de Peer, Y., 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8), pp.2898-2903.

Doronina, L., Churakov, G., Shi, J., Brosius, J., Baertsch, R., Clawson, H. and Schmitz, J., 2015. Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria,

Carnivora). *Molecular Biology and Evolution*, 32(12), pp.3194-3204.

Dugas, D.V., Hernandez, D., Koenen, E.J., Schwarz, E., Straub, S., Hughes, C.E., Jansen, R.K., Nageswara-Rao, M., Staats, M., Trujillo, J.T. and Hajrah, N.H., 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. *Scientific Reports*, 5, p.16958.

Du Puy, D. ed., 2002. The leguminosae of Madagascar. Royal Botanic Gardens Kew.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. and Chen, Z., 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7), p.644.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M. and MacManes, M.D., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), p.1494.

Huang, S., Wu, W., Chen, Z., Zhu, Q., Ng, W.L. and Zhou, Q., 2018. Characterization of the chloroplast genome of *Erythrophleum fordii* (Fabaceae). *Conservation Genetics Resources*, pp.1-3.

Hughes, C.E., Bailey, C.D., Krosnick, S. and Luckow, M.A., 2003. Relationships among genera of the informal Dichrostachys and Leucaena groups (Mimosoideae) inferred from nuclear ribosomal ITS sequences. In: B.B. Klitgaard and A. Bruneau (editors). Advances in Legume Systematics, part 10, Higher Level Systematics, pp. 221–238. Royal Botanic Gardens, Kew.

Huson, D.H., 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1), pp.68-73.

Iganci, J.R., Soares, M.V., Guerra, E. and Morim, M.P., 2016. A preliminary molecular phylogeny of the Abarema alliance (Leguminosae) and implications for taxonomic rearrangement. *International Journal of Plant Sciences*, 177(1), pp.34-43.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J. and Wickett, N.J., 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment.

Applications in Plant Sciences, 4(7), p.1600016.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T. and Leebens-Mack, J.H., 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*, 68(4), pp.594-606.

Jones, M.R. and Good, J.M., 2016. Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25(1), pp.185-202.

Katoh, K., Kuma, K.I., Toh, H. and Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2), pp.511-518.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4), pp.656-664.

Kobert, K., Salichos, L., Rokas, A. and Stamatakis, A., 2016. Computing the internode certainty and related measures from partial gene trees. *Molecular Biology and Evolution*, 33(6), pp.1606-1617.

Kostermans, A.J.G.H., 1954. A Monograph of the Asiatic, Malaysian, Australian, and Pacific Species of Mimosaceae, Formerly Included in Pithecolobium Mart. Organization for Scientific Research in Indonesia.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M. and Karthikeyan, A.S., 2011. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1), pp.D1202-D1210.

Lartillot, N., Rodrigue, N., Stubbs, D. and Richer, J., 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, 62(4), pp.611-615.

Lavin, M., Herendeen, P.S. and Wojciechowski, M.F., 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology*, 54(4), pp.575-594.

Le, S.Q., Lartillot, N. and Gascuel, O., 2008. Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512), pp.3965-3976.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S.O., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R. and Martienssen, R.A., 2011. A functional phylogenomic view of the seed plants. PLoS Genetics, 7(12), p.e1002411.

Lemmon, E.M. and Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44, pp.99-121.

Lewis, G.P., 2005. Legumes of the World. Royal Botanic Gardens Kew.

Li, W. and Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), pp.1658-1659.

Luckow, M., Miller, J.T., Murphy, D.J. and Livshultz, T. 2003. A phylogenetic analysis of the Mimosoideae (Leguminosae) based on chloroplast DNA sequence data. In: B.B. Klitgaard and A. Bruneau (editors). Advances in Legume Systematics, part 10, Higher Level Systematics, pp. 197–220. Royal Botanic Gardens, Kew.

Luckow, M., Fortunato, R.H., Sede, S. and Livshultz, T., 2005. The phylogenetic affinities of two mysterious monotypic mimosoids from southern South America. *Systematic Botany*, 30(3), pp.585-602.

Maddison, W.P., 1997. Gene trees in species trees. *Systematic Biology*, 46(3), pp.523-536.

Magee, A.M., Aspinall, S., Rice, D.W., Cusack, B.P., Sémon, M., Perry, A.S., Stefanović, S., Milbourne, D., Barth, S., Palmer, J.D. and Gray, J.C., 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research*, 20(12), pp.1700-1710.

Mai, U. and Mirarab, S., 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, 19(5), p.272.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H. and Burke, J.M., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences*, 2(2), p.1300085.

Marcet-Houben, M. and Gabaldón, T., 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology*, 13(8), p.e1002220.

Meier, J.I., Marques, D.A., Mwaiko, S., Wagner, C.E., Excoffier, L. and Seehausen, O., 2017.

Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, 8, p.14363.

Miller, J.T., Grimes, J.W., Murphy, D.J., Bayer, R.J. and Ladiges, P.Y., 2003. A phylogenetic analysis of the Acacieae and Ingeae (Mimosoideae: Fabaceae) based on trnK, matK, psbA-trnH, and trnL/trnF sequence data. *Systematic Botany*, 28(3), pp.558-567.

Miller, J.T., Murphy, D.J., Ho, S.Y., Cantrill, D.J. and Seigler, D., 2013. Comparative dating of *Acacia*: combining fossils and multiple phylogenies to infer ages of clades with poor fossil records. *Australian Journal of Botany*, 61(6), pp.436-445.

Moore, A.J., Vos, J.M.D., Hancock, L.P., Goolsby, E. and Edwards, E.J., 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the portullugo clade (Caryophyllales). *Systematic Biology*, 67(3), pp.367-383.

Nicholls, J.A., Pennington, R.T., Koenen, E.J., Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N. and Kidner, C.A., 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science*, 6, p.710.

Ojeda, D.I., Koenen, E., Cervantes, S., de la Estrella, M., Banguera-Hinestroza, E., Janssens, S.B., Migliore, J., Demenou, B.B., Bruneau, A., Forest, F. and Hardy, O.J., 2019. Phylogenomic analyses reveal an exceptionally high number of evolutionary shifts in a florally diverse clade of African legumes. *Molecular Phylogenetics and Evolution*, *137*, pp.156-167.

Pennington, R.T., Prado, D.E. and Pendry, C.A., 2000. Neotropical seasonally dry forests and Quaternary vegetation changes. *Journal of Biogeography*, pp.261-273.

Pennington, R.T., Lavin, M. and Oliveira-Filho, A., 2009. Woody plant diversity, evolution, and ecology in the tropics: perspectives from seasonally dry tropical forests. *Annual Review of Ecology, Evolution, and Systematics*, 40, pp.437-457.

Ranwez, V., Douzery, E.J., Cambon, C., Chantret, N. and Delsuc, F., 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, 35(10), pp.2582-2584.

Ribeiro, P.G., Luckow, M., Lewis, G.P., Simon, M.F., Cardoso, D., de Souza, É.R., Conceicao

Silva, A.P., Jesus, M.C., dos Santos, F.A., Azevedo, V. and de Queiroz, L.P., 2018. *Lachesiodendron*, a new monospecific genus segregated from *Piptadenia* (Leguminosae: Caesalpinioideae: mimosoid clade): Evidence from morphology and molecules. *Taxon*, 67(1), pp.37-54.

Richards, E.J., Brown, J.M., Barley, A.J., Chong, R.A. and Thomson, R.C., 2018. Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic Biology*, 67(5), pp.847-860.

Salichos, L. and Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449), p.327.

Sass, C., Iles, W.J., Barrett, C.F., Smith, S.Y. and Specht, C.D., 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ*, 4, p.e1584.

Sayyari, E. and Mirarab, S., 2018. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes*, 9(3), p.132.

Sayyari, E., Whitfield, J.B. and Mirarab, S., 2018. DiscoVista: Interpretable visualizations of gene tree discordance. Molecular Phylogenetics and Evolution, 122, pp.110-115.

Schmieder, R. and Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), pp.863-864.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J. and Xu, D., 2010. Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278), p.178.

Scornavacca, C. and Galtier, N., 2017. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1), pp.112-120.

Shen, X.X., Hittinger, C.T. and Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5), p.0126.

Smith, S.A., Moore, M.J., Brown, J.W. and Yang, Y., 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15(1), p.150.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312-1313.

Suh, A., Smeds, L. and Ellegren, H., 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology*, 13(8), p.e1002224.

Suh, A., 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zoologica Scripta*, 45, pp.50-62.

Vatanparast, M., Powell, A., Doyle, J.J. and Egan, A.N., 2018. Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences*, 6(3), p.e1036.

Walker, J.F., Brown, J.W. and Smith, S.A., 2018. Analysing contentious relationships and outlier genes in phylogenomics. *Systematic Biology*, 67(5), pp.916-924.

Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. and Liston, A., 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, 2(9), p.1400042.

Wen, J., Xiong, Z., Nie, Z.L., Mao, L., Zhu, Y., Kan, X.Z., Ickert-Bond, S.M., Gerrath, J., Zimmer, E.A. and Fang, X.D., 2013. Transcriptome sequences resolve deep relationships of the grape family. PloS one, 8(9), p.e74394.

Williams, A.V., Boykin, L.M., Howell, K.A., Nevill, P.G. and Small, I., 2015. The complete sequence of the *Acacia ligulata* chloroplast genome reveals a highly divergent clpP1 gene. *PLoS One*, 10(5), p.e0125768.

Wu, F., Mueller, L.A., Crouzillat, D., Pétiard, V. and Tanksley, S.D., 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics*, 174(3), pp.1407-1420.

Yang, Y. and Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, 31(11), pp.3081-3092.

Yang, Y., Moore, M.J., Brockington, S.F., Soltis, D.E., Wong, G.K.S., Carpenter, E.J., Zhang, Y., Chen, L., Yan, Z., Xie, Y. and Sage, R.F., 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Molecular Biology and Evolution, 32(8), pp.2001-2014.

Young, N.D., Debellé, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito,

V.A., Mayer, K.F., Gouzy, J., Schoof, H. and Van de Peer, Y., 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378), p.520.

Zeng, L., Zhang, N., Zhang, Q., Endress, P.K., Huang, J. and Ma, H., 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytologist, 214(3), pp.1338-1354.

Zhang, J., Kobert, K., Flouri, T. and Stamatakis, A., 2013. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5), pp.614-620.

Zhang, C., Rabiee, M., Sayyari, E. and Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6), p.153.

Supplementary Information (see Appendix V on page 266)

Table S1. Voucher details, repository accession numbers and sequencing results for the 122 accessions used in this study.

Figure S1. ML tree of the concatenated amino acid alignment of the 510 gene alignments with more than half of the accessions present, inferred with the LG4X model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S2. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S3. ML phylogeny of 72 protein coding genes from the chloroplast genome inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S4. ML topology of concatenated alignment of 1,767 gene alignments, with ICA values indicated as branch labels. These values were calculated from gene trees of the same 1,767 gene alignments (except with short sequences removed per gene alignment), taking only those bipartitions that received at least 80% BS into account.

Figure S5. ML topology of the concatenated alignment of the 510 gene alignments with more than half of the accessions present, with number of concordant and conflicting gene trees from the same set of 510 alignments written above and below internodes, respectively. Pie charts show the number of concordant bipartitions in blue, the most common conflicting bipartition in green, all other conflicting bipartitions in red and non-informative gene trees in grey. Only bipartitions with at least 50% BS were taken into account.

Figure S6. ASTRAL tree with polytomy test results indicated, only showing non-zero p-values, for nodes with a p-value >0.05 (shown in red) a polytomy is not rejected. Terminal branch lengths are set at 1 (instead of 0) for better visualization.

Figure S7. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, but with Cedrelinga cateniformis removed, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S8. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, but with Cedrelinga cateniformis and the Samanea clade removed, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

CHAPTER III

Conclusions

Phylogenetic tree-thinking has long permeated evolutionary biology and currently lies very much at the heart of phylogenomics, even though the Tree of Life metaphor has been challenged by the widespread occurrence of incomplete lineage sorting (ILS), introgression/hybridization and lateral gene transfer (LGT). This thesis shows that several of these issues and challenges are also prominent in the legume family, where organismal evolution has been highly complex at the genomic level, hindering the reconstruction of the most likely species tree and suggesting that there are significant hard polytomies, as well as reticulation, deeply embedded in the legume phylogeny.

In Chapter I of this thesis, I have shown that the most likely relationships among legume subfamilies, in other words the earliest dichotomies in the legume phylogeny, are only supported by a slightly higher fraction of gene trees than alternative topologies, while at the same time, many gene trees show a lack of phylogenetic signal for this part of the phylogeny. While part of this observed gene tree conflict may be caused by gene tree estimation errors, extensive ILS has likely occurred. This conflict is probably a common feature of eukaryotic organismal evolution more generally, particularly when lineage diversification occurs rapidly as has likely been the case in the initial radiation of the legumes into six major lineages.

The occurrence of multiple paleopolyploidy events (or ancient whole genome duplications or WGDs) early in the evolution of the legumes, including one likely allopolyploid event, has further complicated the reconstruction of genome evolution across the deepest divergences in the family, as shown in Chapter II. Taken together, these results suggest that the early evolution of the legumes is best represented by a network, which better reconciles the evolutionary history of different genes and other genomic elements than a strictly bifurcating phylogeny, although ILS still needs to be invoked as well.

Furthermore, I demonstrate in Chapter II that the timings of the initial diversification of the legumes and of ancient polyploidy events are close to 66 Ma at the Cretaceous-Paleogene boundary (KPB). This suggests that the opportunities offered by large-scale evolutionary turnover caused by the KPB mass extinction event and the resulting emergence of new habitats, as well as the expanded genomic substrate following multiple polyploidization

events, can together contribute to explain the spectacular evolutionary success of the legumes in the Cenozoic.

Finally, in Chapter III I present an analysis of a more recent episode of rapid diversification within the mimosoid clade of legumes. Apart from several nodes with gene tree conflict involving three lineages, a near-complete lack of resolution is observed across all 1,767 analysed gene trees involving the relationships of 6 or 7 subclades within the Ingioid clade, suggesting this constitutes a single large hard polytomy.

These results are highly significant for our understanding of the evolution of the legume family and of angiosperms more generally. All three chapters highlight the complexity of genome evolution in deep time where several different processes and events such as ILS, polyploidy, hybridization and rapid lineage diversification cause problems for inferring bifurcating species trees. This adds to a growing number of phylogenomic studies of various taxonomic groups, including in angiosperms, that show how evolution at the level of genomes and organisms is far from always tree-like. The reticulate polyploid evolution and hard polytomies revealed here in legumes are probably not in any way exceptional across angiosperms and other organismal groups, as many ongoing and new phylogenomics research projects in the near future may well demonstrate. The timings of rapid lineage diversification and WGDs are also relevant for understanding the role of geological history in shaping biodiversity. While prominent animal clades have been particularly well-studied in relation to turnover at mass extinction events, I have demonstrated that much like mammals and birds, the legumes also rapidly diversified in the early Cenozoic following the KPB. Similarly, the radiation of the Ingioid clade likely occurred during the transition from the hothouse climate of the Early Eocene to the much cooler Miocene. In both cases, there appear to be links between geological events that led to the emergence of new habitats and the observed patterns of genome evolution brought about for example by selection favouring polyploids and their associated traits, and rapid lineage diversification leading to lack of phylogenetic signal and ILS.

While networks and/or hard polytomies are gradually coming to be seen as better representations of evolutionary history than a strictly bifurcating tree for many groups, inferring a species tree using either concatenation or gene tree summarisation approaches is

still commonplace. The main drawback of gene tree approaches is that single genes are often not informative enough to resolve relationships, especially if taxa rapidly diversified. For larger phylogenies with dense sampling of species, the proportion of conflicting or uninformative gene trees for individual nodes is still expected to be relatively high due to ILS and larger numbers of speciation events. On the other hand, analyses of concatenated phylogenomic data often find a fully resolved tree, but support values for potentially erroneous species relationships may be inflated in these analyses. Gene jackknifing may be viewed as a useful 'intermediate' option and presents one of the best methods currently available to evaluate which nodes we can and cannot resolve given the available data.

The rapid accumulation of genome-scale data across taxonomic groups is set to continue and indeed rapidly accelerate further in the next decade. This forthcoming proliferation of phylogenomic data will offer exciting new research possibilities and promises to deliver many interesting new discoveries. Rather than ending phylogenetic incongruence, phylogenomics has led to a much more comprehensive understanding of the causes of incongruence and the molecular aspects of organismal evolution in general. While many studies still focus on resolving relationships, as I have done in Chapters I and III, much can be learned from studying the interactions among genomes and the environment in shaping molecular and organismal evolution. Several studies have already started to address these questions (e.g. Brawand et al., 2014, Lamichhaney et al., 2015, Pease et al., 2016, Moore et al., 2017), and large-scale studies of functional evolution in phylogenomics are set to be one of the main directions for systematic biology in the near future. For example, gene duplications and losses may have led to gene family expansions or reductions related to their function and environmental change (Griessman et al., 2018). Similarly, a proliferation of polyploid lineages as has been suggested to be associated with the mass extinction event at the KPB, could occur in various other environmental settings or may be related to environmental change more generally (Cai et al., 2019). Moreover, the role of selection at the level of amino acids is likely important for local adaptation to different environmental factors and for speciation, but functional evolution of protein-coding genes remains poorly understood in relation to geological history and the origination of diverse life-history strategies (Lee et al., 2011). These and numerous other topics can be tackled in phylogenomics when large

numbers of high-quality genome assemblies are available across groups of interest and when inter- and infraspecific genetic variation are characterized by sequencing multiple individuals when studying closely related species. The field of phylogenomics is therefore likely to continue to flourish as ongoing technological developments and increased sequencing efforts and efficiencies lead to ever larger data sets.

Now that phylogenomics has overcome the limitations of sparse gene sampling, it seems timely to shift the focus to sampling taxa more densely. Greatly expanded taxon sampling is likely to be equally (or even more) beneficial than further increasing the volume of sequence data because, with denser taxon sampling, orthology can be more accurately assessed, and both gene tree and species tree analyses will be more accurate and less prone to LBA artefacts. Phylogenomic analyses with dense taxon sampling, coupled with careful evaluation of conflicting phylogenetic signals, presents a powerful way forward in systematics and evolutionary ecology, especially for diverse and relatively recently evolved clades. A great deal of recent work has focused on resolving deep divergences with phylogenomic data sets. While this is understandable given the sparse availability of genomic data across the Tree of Life until recently, with modern techniques we are now able to apply genome-scale data analysis to diverse clades with dense taxon sampling to solve the myriad of taxonomic problems that exist mainly at the levels of tribes and genera, and for which traditional chloroplast and nuclear ribosomal markers have proved to be insufficiently informative.

Sequencing of many more legume genomes, including *Duparquetia* - the sole member of Duparquetioideae for which genome-scale nuclear data are as yet unavailable - alongside key taxa in other subfamilies, is planned over the next few years as part of the 10KP initiative (Cheng et al., 2019). Furthermore, hybrid capture sequencing projects are planned or ongoing in all non-monotypic legume subfamilies (Detarioideae, de la Estrella et al., unpublished data; Cercidoideae and Dialioideae, Bruneau et al., unpublished data; Caesalpinioideae, Koenen et al., Chapter III and Ringelberg & Koenen et al., unpublished data; Papilionoideae, Vatanparast & Egan et al., unpublished data). In addition, as part of the Plant and Fungal Tree of Life (PAFTOL) project of the Royal Botanic Gardens at Kew, every legume genus is planned to be sequenced using hybrid capture of a universal set of 353 angiosperm genes

(Johnson et al., 2018). Analyses of all of these data will undoubtedly lead to better and more comprehensive estimates of the legume phylogeny. However, combining all these data sets to build a legume phylogeny at the same taxonomic scale as the *matK* tree of LPWG (2017) is a potentially daunting task. As a way forward, I suggest that it is unfeasible, and indeed unnecessary, to attempt to create a single supermatrix for several hundreds or thousands of taxa and hundreds of genes, but that instead it would be more sensible to infer a backbone phylogeny using complete genomes (such as the one presented in Chapter I but with extended taxon sampling) alongside independent phylogenies based on large (but not necessarily identical gene sets) for each subfamily, which can then be combined using a tree-grafting approach that I describe further below. This approach ensures that resolution along the backbone or within each subfamily or subclade is not compromised by poor overlap among genes, or paralogy issues among subfamilies, while at the same time, the analyses remain computationally tractable.

Apart from enhancing the legume phylogeny, completely sequenced and well-annotated genomes for each of the subfamilies can be used to answer many questions about early legume genome evolution and polyploidy and the evolutionary origins of legume traits. The study of paleopolyploidy in legumes would be greatly advanced by the availability of high quality complete genomes of all subfamilies, in particular using the positional homology data that can be harvested from these. Furthermore, other types of marker loci, such as retroposon insertion sites which are homoplasy-free (Suh et al., 2015), can also be extracted from high-quality genome assemblies to quantify the strength of ILS, detect introgression or more accurately reconstruct (allo)polyploid histories. Undoubtedly, the evolution of nodulation in legumes will remain an important research theme, but also the evolution of genes underlying various other traits, such as floral symmetry and fruit dehiscence, could be studied using phylogenomic methods.

For the mimosoid clade, expansion of the hybrid capture data set that is presented in Chapter III to c. 450 accessions across Caesalpinioideae as a whole is currently underway in the lab in Zurich. Using these data, we intend to infer an enhanced backbone phylogeny for subfamily Caesalpinioideae to be used, in combination with species-level data sets, for large scale biogeographic and macro-evolutionary studies. I have carried out preliminary analyses

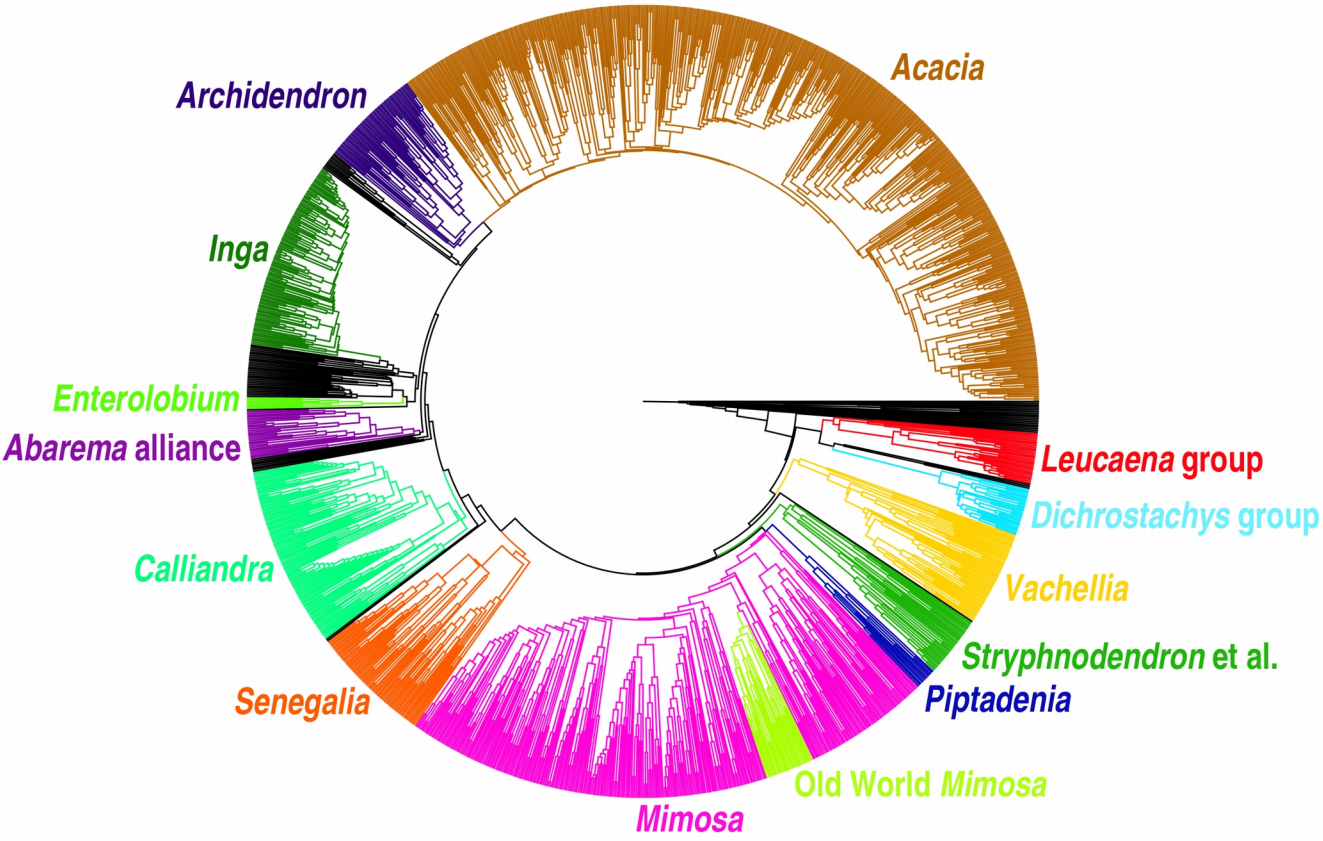


Figure 1. Mimosoid mega-tree indicating the grafted subtrees, with the hybrid capture backbone shown in black. There are 1,452 species represented in the tree, c. 44% of the total number of species in the mimosoid clade.

for the mimosoid clade, where the hybrid capture backbone from Chapter III was used as a rootstock to 'graft' a set of 14 densely sampled phylogenetic trees for genera or groups of closely related genera onto (Fig. 1), similar to how Spriggs et al. (2014) built a large phylogenetic tree for the grasses. The resulting mega-tree is a chronogram, and it shows that core mimosoids originated at c. 35 Ma, while the initial radiation of the Ingioid clade is estimated to have occurred between c. 30-25 Ma, suggesting that mimosoids diversified mainly in response to the cooling and drying trends that started in the late Eocene and which characterizes the whole of the Oligocene (Fig. 2). Basic mapping of biogeography and biome associations using this mega-tree further show that mimosoids originated in the Early Eocene wet forests of Africa, with several species-poor African wet forest lineages forming successive

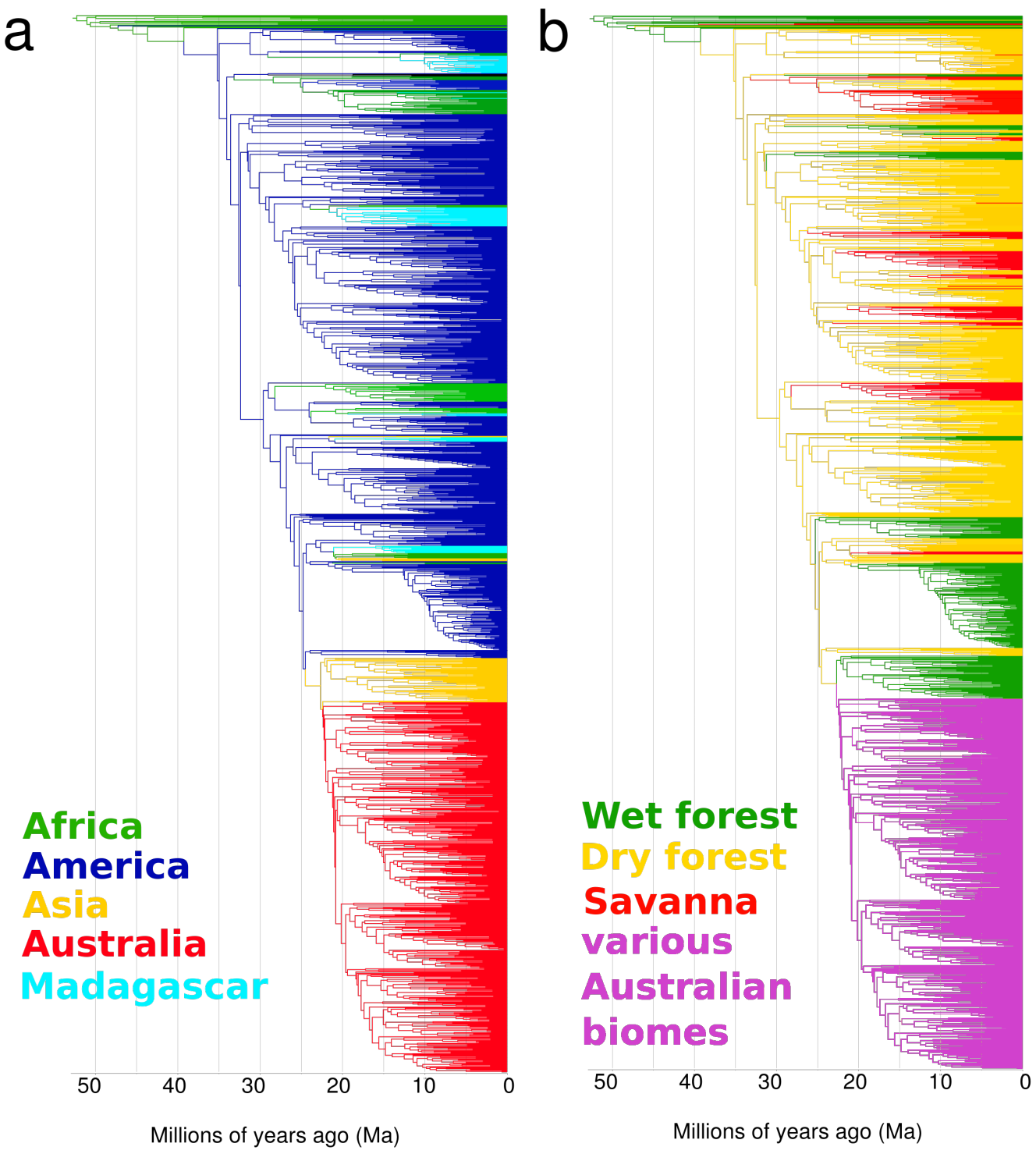


Figure 2. Mimosoid mega-tree with (a) biogeography and (b) biome association indicated as per the legend.

sister-groups (or a paraphyletic grade) to the core mimosoids. With the exception of 1000+ species in the Australasian clade of Ingioids including *Acacia* s.s., the core mimosoids predominantly occur in the American tropics and mainly occur in seasonally dry vegetation (Fig. 2). Embedded within the core mimosoids, which are presumably ancestrally adapted to the dry tropics, are at least six clades that have switched back to the rainforest biome and at least eight clades have switched to the savanna biome (Fig. 2b).

These results are preliminary, and more carefully elaborated analyses are necessary, but nevertheless illustrate the potential of having large and well-resolved phylogenetic trees available for large scale comparative analysis of clades of several thousands of species. These types of analyses, coupled with estimation of speciation rates across the tree and investigations into the role of extinction, will be used with the expanded phylogenomic backbone data set to test hypotheses about diversification of lowland tropical vegetation in relation to past environmental change, for which the Caesalpinioideae form an excellent study group. Apart from this, the expanded hybrid capture phylogeny will also be used to establish a new tribal and clade-based classification for Caesalpinioideae and to aid generic re-delimitation in the mimosoid clade. Finally, the genus *Albizia* is nearly completely sampled in the expanded hybrid capture data set, such that a species-level phylogeny for the genus can be constructed and used for taxonomic revision and to study ecological diversification. Species of *Albizia* in Africa, Madagascar and Asia have adapted to nearly all tropical vegetation types including the Congo basin lowland rainforests, the Kalahari and Namib deserts, Albertine rift montane forests, seasonally dry tropical forests and spiny scrub in Madagascar and the Horn of Africa, continental Asian subtropical upland forests and Malesian monsoonal vegetation, making *Albizia* an ideal study system to investigate ecological radiation.

To conclude, this thesis significantly enhances our knowledge about the early evolution of the legumes, the phylogenetic relationships among legume subfamilies and within the mimosoid clade, and our understanding of the complexity of genome evolution in deep time and its consequences for inferring the Tree of Life. Furthermore, it signals in which directions the study of legume paleopolyploidy and ancestral genome reconstruction may proceed and provides a basis for further macro-evolutionary and biogeographical studies using the

Caesalpinioideae as a model clade for lowland tropical biodiversity. More generally, this thesis forms an important contribution to the study of genome evolution in angiosperms in relation to the origins of Cenozoic biodiversity and manifests the limits to phylogenetic resolution across the eukaryotic Tree of Life, caused by rapid diversification, polyploidization, ILS, hybridization and/or methodological limitations. As I have shown in this thesis, all of the above may likely play a role in causing lack of resolution in portions of the legume phylogeny.

References

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E. and Turner-Maier, J., 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):375.

Cai, L., Xi, Z., Amorim, A.M., Sugumaran, M., Rest, J.S., Liu, L. and Davis, C.C., 2019. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytologist*, 221: 565-576.

Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.-M., Li F.-W., Melkonian B., Mavrodiev E.V., Sun W., Fu Y., Yang H., Soltis D.E., Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K.-S. 2018. 10KP: a phylodiverse genome sequencing plan. GigaScience. 7:1–9.

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B., Billault-Penneteau, B., Lauressergues, D., Keller, J., Imanishi, L. and Roswanjaya, Y.P., 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*, 361(6398):eaat1743.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T. and Leebens-Mack, J.H., 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*, 68(4):594-606.

Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C.J., Wang, C., Zamani, N. and Grant, B.R., 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539):371.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R., Martienssen, R.A., Coruzzi, G. & DeSalle, R., 2011. A Functional Phylogenomic View of the Seed Plants. *PLoS Genetics*, 7: e1002411.

LPWG (Legume Phylogeny Working Group), 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon*, 66:44–77.

Moore, A.J., Vos, J.M.D., Hancock, L.P., Goolsby, E. and Edwards, E.J., 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the portullugo clade (Caryophyllales). *Systematic Biology*, 67(3):367-383.

Pease, J.B., Haak, D.C., Hahn, M.W. and Moyle, L.C., 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology*, 14(2):e1002379.

Spriggs, E.L., Christin, P.A. and Edwards, E.J., 2014. C4 photosynthesis promoted species diversification during the Miocene grassland expansion. *PloS ONE*, 9(5), p.e97722.

Suh, A., Smeds, L., Ellegren, H., 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology* 13(8):e1002224.

Appendix I Abstracts of co-authored publications

Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species–rich clades

Legume Phylogeny Working Group, Bruneau, A., Doyle, J. J., Herendeen, P., Hughes, C., Kenicer, G., Lewis, G., Mackinder, B., Pennington, R. T., Sanderson, M. J., Wojciechowski, M. F., Boatwright, S., Brown, G., Cardoso, D., Crisp, M., Egan, A., Fortunato, R. H., Hawkins, J., Kajita, T., Klitgaard, B., **Koenen, E.**, Lavin, M., Luckow, M., Marazzi, B., McMahon, M. M., Miller, J. T., Murphy, D. J., Ohashi, H., de Queiroz, L. P., Rico, L., Särkinen, T., Schrire, B., Simon, M. F., Souza, E. R., Steele, K., Torke, B. M., Wieringa, J. J. and van Wyk, B. – *Taxon* (2013) 62:217-248.

The Leguminosae, the third–largest angiosperm family, has a global distribution and high ecological and economic impor tance. We examine how the legume systematic research community might join forces to produce a comprehensive phylogenetic estimate for the ca. 751 genera and ca. 19,500 species of legumes and then translate it into a phylogeny–based classification. We review the current state of knowledge of legume phylogeny and highlight where problems lie, for example in taxon sampling and phylogenetic resolution. We review approaches from bioinformatics and next–generation sequencing, which can facilitate the production of better phylogenetic estimates. Finally, we examine how morphology can be incorporated into legume phylogeny to address issues in comparative biology and classification. Our goal is to stimulate the research needed to improve our knowledge of legume phylogeny and evolution; the approaches that we discuss may also be relevant to other species-rich angiosperm clades.

Keywords: Caesalpinioideae, Fabaceae, Leguminosae, low-copy nuclear genes, Mimosoideae, multiple sequence alignment, Papilionoideae, phylogenetic inference

My contributions: I co-wrote the section on mimosoid legumes.

APPENDIX I

Exploring the tempo of species diversification in legumes

Koenen, E.J.M.*, De Vos, J.M.*, Atchison, G.W., Simon, M.F., Schrire, B.D., De Souza, E.R., de Queiroz, L.P. and Hughes, C.E. – *South African Journal of Botany* (2013) 89*:*19-30.

Whatever criteria are used to measure evolutionary success – species numbers, geographic range, ecological abundance, ecological and life history diversity, background diversification rates, or the presence of rapidly evolving clades – the legume family is one of the most successful lineages of flowering plants. Despite this, we still know rather little about the dynamics of lineage and species diversification across the family through the Cenozoic, or about the underlying drivers of diversification. There have been few attempts to estimate net species diversification rates or underlying speciation and extinction rates for legume clades, to test whether among-lineage variation in diversification rates deviates from null expectations, or to locate species diversification rate shifts on specific branches of the legume phylogenetic tree. In this study, time-calibrated phylogenetic trees for a set of species-rich legume clades – *Calliandra*, Indigofereae, *Lupinus*, *Mimosa* and Robinieae – and for the legume family as a whole, are used to explore how we might approach these questions. These clades are analysed using recently developed maximum likelihood and Bayesian methods to detect species diversification rate shifts and test for among-lineage variation in speciation, extinction and net diversification rates. Possible explanations for rate shifts in terms of extrinsic factors and/or intrinsic trait evolution are discussed. In addition, several methodological issues and limitations associated with these analyses are highlighted emphasizing the potential to improve our understanding of the evolutionary dynamics of legume diversification by using much more densely sampled phylogenetic trees that integrate information across broad taxonomic, geographical and temporal levels.

Keywords: Species diversification, Leguminosae, *Calliandra*, Indigofereae, *Lupinus*, *Mimosa*, Robinieae, Diversification rate shift, Speciation rate, Extinction rate.

* these authors contributed equally to this work.

My contributions: I carried out part of the analyses, designed 2 of the 3 figures and co-wrote the paper.

Towards a new classification system for legumes: Progress report from the 6th International Legume Conference

The Legume Phylogeny Working Group, Borges, L., Bruneau, A., Cardoso, D., Crisp, M., Delgado-Salinas, A., Doyle, J.J., Egan, A., Herendeen, P.S., Hughes, C., Kenicer, G., Klitgaard, B., **Koenen, E.**, Lavin, M., Lewis, G., Luckow, M., Mackinder, B., Malécot, V., Miller, J.T., Pennington, R.T., de Queiroz, L.P., Schrire, B., Simon, M.F., Steele, K., Torke, B., Wieringa, J.J., Wojciechowski, M.F., Boatwright, S., de la Estrella, M., de Freitas Mansano, V., Prado, D.E., Stirton, C., Wink, M. – *South African Journal of Botany* (2013) 89:3-9.

Legume systematists have been making great progress in understanding evolutionary relationships within the Leguminosae (Fabaceae), the third largest family of flowering plants. As the phylogenetic picture has become clearer, so too has the need for a revised classification of the family. The organization of the family into three subfamilies and 42 tribes is outdated and evolutionarily misleading. The three traditionally recognized subfamilies, Caesalpinioideae, Mimosoideae, and Papilionoideae, do not adequately represent relationships within the family. The occasion of the Sixth International Legume Conference in Johannesburg, South Africa in January 2013, with its theme “Towards a new classification system for legumes,” provided the impetus to move forward with developing a new classification. A draft classification, based on current phylogenetic results and a set of principles and guidelines, was prepared in advance of the conference as the basis for discussion. The principles, guidelines, and draft classification were presented and debated at the conference. The objectives of the discussion were to develop consensus on the principles that should guide the development of the classification, to discuss the draft classification's strengths and weaknesses and make proposals for its revision, and identify and prioritize phylogenetic deficiencies that must be resolved before the classification could be published. This paper describes the collaborative process by a large group of legume systematists, publishing under the name Legume Phylogeny Working Group, to develop a new phylogenetic classification system for the Leguminosae. The goals of this paper are to inform the broader legume community, and others, of the need for a revised classification, and spell out clearly what the alternatives and challenges are for a new classification system for the family.

Keywords: Leguminosae, Classification, Fabaceae

My contributions: I co-wrote the aricle.

APPENDIX I

Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae)

Nicholls, J.A., Pennington, R.T., **Koenen, E.J.**, Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N. and Kidner, C.A. – *Frontiers in Plant Science* (2015) 6:710.

Evolutionary radiations are prominent and pervasive across many plant lineages in diverse geographical and ecological settings; in neotropical rainforests there is growing evidence suggesting that a significant fraction of species richness is the result of recent radiations. Understanding the evolutionary trajectories and mechanisms underlying these radiations demands much greater phylogenetic resolution than is currently available for these groups. The neotropical tree genus *Inga* (Leguminosae) is a good example, with ~300 extant species and a crown age of 2–10 MY, yet over 6 kb of plastid and nuclear DNA sequence data gives only poor phylogenetic resolution among species. Here we explore the use of larger-scale nuclear gene data obtained though targeted enrichment to increase phylogenetic resolution within *Inga*. Transcriptome data from three *Inga* species were used to select 264 nuclear loci for targeted enrichment and sequencing. Following quality control to remove probable paralogs from these sequence data, the final dataset comprised 259,313 bases from 194 loci for 24 accessions representing 22 *Inga* species and an outgroup (*Zygia*). Bayesian phylogenies reconstructed using either all loci concatenated or a gene-tree/species-tree approach yielded highly resolved phylogenies. We used coalescent approaches to show that the same targeted enrichment data also have significant power to discriminate among alternative within-species population histories within the widespread species *I. umbellifera*. In either application, targeted enrichment simplifies the informatics challenge of identifying orthologous loci associated with *de novo* genome sequencing. We conclude that targeted enrichment provides the large volumes of phylogenetically-informative sequence data required to resolve relationships within recent plant species radiations, both at the species level and for within-species phylogeographic studies.

Keywords: hybrid capture, Inga, next-generation sequencing, phylogenomics, population genomics, radiation, targeted enrichment

My contributions: This article is the result of a collaboration on hybrid capture in mimosoids, where we designed baits together for studies in *Inga* and across the whole of the mimosoid clade. Apart from having been involved in the bait design, I have provided extensive comments on the analyses and contributed to the writing of the article.

Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*

Dugas, D.V., Hernandez, D., **Koenen, E.J.**, Schwarz, E., Straub, S., Hughes, C.E., Jansen, R.K., Nageswara-Rao, M., Staats, M., Trujillo, J.T. and Hajrah, N.H. – *Scientific reports* (2015) 5:16958.

The Leguminosae has emerged as a model for studying angiosperm plastome evolution because of its striking diversity of structural rearrangements and sequence variation. However, most of what is known about legume plastomes comes from few genera representing a subset of lineages in subfamily Papilionoideae. We investigate plastome evolution in subfamily Mimosoideae based on two newly sequenced plastomes (*Inga* and *Leucaena*) and two recently published plastomes (*Acacia* and *Prosopis*), and discuss the results in the context of other legume and rosid plastid genomes. Mimosoid plastomes have a typical angiosperm gene content and general organization as well as a generally slow rate of protein coding gene evolution, but they are the largest known among legumes. The increased length results from tandem repeat expansions and an unusual 13 kb IR-SSC boundary shift in *Acacia* and *Inga*. Mimosoid plastomes harbor additional interesting features, including loss of *clpP* intron1 in *Inga*, accelerated rates of evolution in *clpP* for *Acacia* and *Inga*, and *dN/dS* ratios consistent with neutral and positive selection for several genes. These new plastomes and results provide important resources for legume comparative genomics, plant breeding, and plastid genetic engineering, while shedding further light on the complexity of plastome evolution in legumes and angiosperms.

My contributions: I have assembled the *Inga* plastome and discovered the boundary shift of the inverted repeat with the short single copy region which has led to an extended inverted repeat in most Ingioid mimosoids, and I contributed to the writing of the article.

APPENDIX I

A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny

The Legume Phylogeny Working Group (LPWG), Azani, N., Babineau, M., Bailey, C.D., Banks, H., Barbosa, A.R., Pinto, R.B., Boatwright, J.S., Borges, L.M., Brown, G.K., Bruneau, A., Candido, E., Cardoso, D., Chung, K., Clark, R.P., Conceição, A.d., Crisp, M., Cubas, P., Delgado-Salinas, A., Dexter, K.G., Doyle, J.J., Duminil, J., Egan, A.N., de la Estrella, M., Falcão, M.J., Filatov, D.A., Fortuna-Perez, A.P., Fortunato, R.H., Gagnon, E., Gasson, P., Rando, J.G., de Azevedo Tozzi, A.M., Gunn, B., Harris, D., Haston, E., Hawkins, J.A., Herendeen, P.S., Hughes, C.E., Iganci, J.R., Javadi, F., Kanu, S.A., Kazempour-Osaloo, S., Kite, G.C., Klitgaard, B.B., Kochanovski, F.J., **Koenen, E.J.**, Kovar, L., Lavin, M., le Roux, M., Lewis, G.P., de Lima, H.C., López-Roberts, M.C., Mackinder, B., Maia, V.H., Malécot, V., Mansano, V.F., Marazzi, B., Mattapha, S., Miller, J.T., Mitsuyuki, C., Moura, T., Murphy, D.J., Nageswara-Rao, M., Nevado, B., Neves, D., Ojeda, D.I., Pennington, R.T., Prado, D.E., Prenner, G., de Queiroz, L.P., Ramos, G., Filardi, F.L., Ribeiro, P.G., de Lourdes Rico-Arce, M., Sanderson, M.J., Santos-Silva, J., São-Mateus, W.M., Silva, M.J., Simon, M.F., Sinou, C., Snak, C., de Souza, É.R., Sprent, J., Steele, K.P., Steier, J.E., Steeves, R., Stirton, C.H., Tagane, S., Torke, B.M., Toyama, H., da Cruz, D.T., Vatanparast, M., Wieringa, J.J., Wink, M., Wojciechowski, M.F., Yahara, T., Yi, T., Zimmerman, E. – *Taxon* (2017) 66:44-77.

The classification of the legume family proposed here addresses the long-known non-monophyly of the traditionally recognised subfamily Caesalpinioideae, by recognising six robustly supported monophyletic subfamilies. This new classification uses as its framework the most comprehensive phylogenetic analyses of legumes to date, based on plastid matK gene sequences, and including near-complete sampling of genera (698 of the currently recognised 765 genera) and ca. 20% (3696) of known species. The matK gene region has been the most widely sequenced across the legumes, and in most legume lineages, this gene region is sufficiently variable to yield well-supported clades. This analysis resolves the same major clades as in other phylogenies of whole plastid and nuclear gene sets (with much sparser taxon sampling). Our analysis improves upon previous studies that have used large phylogenies of the Leguminosae for addressing evolutionary questions, because it maximises generic sampling and provides a phylogenetic tree that is based on a fully curated set of sequences that are vouchered and taxonomically validated. The phylogenetic trees obtained and the underlying data are available to browse and download, facilitating subsequent analyses that require evolutionary trees. Here we propose a new community-endorsed classification of the family that reflects the phylogenetic structure that is consistently resolved

and recognises six subfamilies in Leguminosae: a recircumscribed Caesalpinioideae DC., Cercidoideae Legume Phylogeny Working Group (stat. nov.), Detarioideae Burmeist., Dialioideae Legume Phylogeny Working Group (stat. nov.), Duparquetioideae Legume Phylogeny Working Group (stat. nov.), and Papilionoideae DC. The traditionally recognised subfamily Mimosoideae is a distinct clade nested within the recircumscribed Caesalpinioideae and is referred to informally as the mimosoid clade pending a forthcoming formal tribal and/or clade-based classification of the new Caesalpinioideae. We provide a key for subfamily identification, descriptions with diagnostic charactertistics for the subfamilies, figures illustrating their floral and fruit diversity, and lists of genera by subfamily. This new classification of Leguminosae represents a consensus view of the international legume systematics community; it invokes both compromise and practicality of use.

Keywords: Caesalpinioideae, Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae, mimosoid clade, Papilionoideae, plastid matK phylogeny

My contributions: I curated and aligned the *matK* data set for mimosoids, combined it with data sets for the rest of the family that were curated by others, and performed the phylogenetic analyses on the resulting matrix. I have also written the formal description of subfamily Caesalpinioideae as well as contributing to the writing of other sections of the article, helped to prepare two figures and contributed photographs for the figures and colour plates.

APPENDIX I

Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes

Couvreur, T.L., Helmstetter, A.J., **Koenen, E.J.**, Bethune, K., Brandão, R.D., Little, S.A., Sauquet, H. and Erkens, R.H. – *Frontiers in plant science* (2019) 9*:*1941.

Targeted enrichment and sequencing of hundreds of nuclear loci for phylogenetic reconstruction is becoming an important tool for plant systematics and evolution. Annonaceae is a major pantropical plant family with 110 genera and ca. 2,450 species, occurring across all major and minor tropical forests of the world. Baits were designed by sequencing the transcriptomes of five species from two of the largest Annonaceae subfamilies. Orthologous loci were identified. The resulting baiting kit was used to reconstruct phylogenetic relationships at two different levels using concatenated and gene tree approaches: a family wide Annonaceae analysis sampling 65 genera and a species level analysis of tribe Piptostigmateae sampling 29 species with multiple individuals per species. DNA extraction was undertaken mainly on silicagel dried leaves, with two samples from herbarium dried leaves. Our kit targets 469 exons (364,653 bp of sequence data), successfully capturing sequences from across Annonaceae. Silicagel dried and herbarium DNA worked equally well. We present for the first time a nuclear gene-based phylogenetic tree at the generic level based on 317 supercontigs. Results mainly confirm previous chloroplast based studies. However, several new relationships are found and discussed. We show significant differences in branch lengths between the two large subfamilies Annonoideae and Malmeoideae. A new tribe, Annickieae, is erected containing a single African genus *Annickia*. We also reconstructed a well-resolved species-level phylogenetic tree of the Piptostigmteae tribe. Our baiting kit is useful for reconstructing well-supported phylogenetic relationships within Annonaceae at different taxonomic levels. The nuclear genome is mainly concordant with plastome information with a few exceptions. Moreover, we find that substitution rate heterogeneity between the two subfamilies is also found within the nuclear compartment, and not just plastomes and ribosomal DNA as previously shown. Our results have implications for understanding the biogeography, molecular dating and evolution of Annonaceae.

Keywords: Annonaceae, rain forests, systematics, transcriptomes, Piptostigmateae, herbarium

My contributions: I used the methodology that was developed for selecting low-copy nuclear genes for hybrid capture in mimosoids to design a set of genes to target in Annonaceae, the results of which are described in this article. I also contributed to the writing of the article.

Phylogenomic analyses reveal an exceptionally high number of evolutionary shifts in a florally diverse clade of African legumes

Ojeda, D.I., **Koenen, E.**, Cervantes, S., de la Estrella, M., Banguera-Hinestroza, E., Janssens, S.B., Migliore, J., Demenou, B.B., Bruneau, A., Forest, F. and Hardy, O.J. – *Molecular phylogenetics and evolution* (2019) 137*:*156-167.

Detarioideae is well known for its high diversity of floral traits, including flower symmetry, number of organs, and petal size and morphology. This diversity has been characterized and studied at higher taxonomic levels, but limited analyses have been performed among closely related genera with contrasting floral traits due to the lack of fully resolved phylogenetic relationships. Here, we used four representative transcriptomes to develop an exome capture (target enrichment) bait for the entire subfamily and applied it to the Anthonotha clade using a complete data set (61 specimens) representing all extant floral diversity. Our phylogenetic analyses recovered congruent topologies using ML and Bayesian methods. *Anthonotha* was recovered as monophyletic contrary to the remaining three genera (*Englerodendron*, *Isomacrolobium* and *Pseudomacrolobium*), which form a monophyletic group sister to *Anthonotha*. We inferred a total of 35 transitions for the seven floral traits (pertaining to flower symmetry, petals, stamens and staminodes) that we analyzed, suggesting that at least 30% of the species in this group display transitions from the ancestral condition reconstructed for the Anthonotha clade. The main transitions were towards a reduction in the number of organs (petals, stamens and staminodes). Despite the high number of transitions, our analyses indicate that the seven characters are evolving independently in these lineages. Petal morphology is the most labile floral trait with a total of seven independent transitions in number and seven independent transitions to modification in petal types. The diverse petal morphology along the dorsoventral axis of symmetry within the flower is not associated with differences at the micromorphology of petal surface, suggesting that in this group all petals within the flower might possess the same petal identity at the molecular level. Our results provide a solid evolutionary framework for further detailed analyses of the molecular basis of petal identity.

Keywords: Detarioideae, Berlinia clade, Flower evolution, Papillose conical cells, Petal number, Petal identity, Phylogenomics, Target enrichment

My contributions: I used the methodology that was developed for selecting low-copy nuclear genes for hybrid capture in mimosoids to design a set of genes to target in Detarioideae, the results of which are described in this article. I also contributed to the writing of the article.

Appendix II Abstracts of talks at conferences

6th International Legume Conference in 2013, Johannesburg, South Africa,
Oral presentation of 15 minutes
Symposium: Phylogeny and classification of Mimosoideae

A novel approach using hybrid sequence capture and massively parallel sequencing to build the Mimosoid Legume phylogeny

E.J.M. Koenen[1], R.T. Pennington[2], C.A. Kidner[2,3], J.A. Nicholls[4], J.T. Miller[5], C.E. Hughes[1]
[1]*Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, 8008 Zürich, Switzerland*
[2]*Royal Botanic Gardens Edinburgh, Edinburgh, UK*
[3]*Institute of Molecular Plant Sciences, University of Edinburgh, UK*
[4]*Institute of Evolutionary Biology, University of Edinburgh, UK*
[5]*Centre for Australian National Biodiversity Research, CSIRO Plant Industry, Canberra, Australia*

The phylogeny of the Legume subfamily Mimosoideae remains poorly resolved, with several large polytomies across the (supra) generic backbone. Lack of resolution is especially stark in the large clade comprising tribe Ingeae plus *Acacia* s.s. that contains almost two thirds of the c. 3271 species in the subfamily. Across this large clade generic relationships are essentially unknown because the sparse resolution in current phylogenies is mostly poorly supported. Moreover, generic delimitation remains in a state of flux, especially surrounding the genus *Albizia* and allies. To address these problems, we are using novel next-generation sequencing approaches to increase the amount of DNA sequence data for more powerful phylogenetic inference in the subfamily. Using hybrid sequence capture techniques with baits produced using PCR, we are sequencing whole chloroplast genomes across mimosoids to resolve the generic backbone. In addition, we have generated RNA-seq data for six species in order to select orthologous single-copy genes from the nuclear genome for further hybrid capture experiments. Recently developed deep multiplexing strategies will be used to greatly upscale the number of samples that can be processed on high-throughput sequencers, and hence taxon sampling. Our goal is to generate a densely sampled phylogeny for the subfamily, with a focus on Ingeae, to resolve higher level relationships and revise the generic delimitation of *Albizia*. The resulting phylogeny will also be used to study large scale diversity dynamics in the tropics throughout the Cenozoic, for which the clade provides an excellent model.

11th Latin American Botanical Congress in 2014, Salvador da Bahia, Brazil
Oral presentation of 15 minutes
Symposium: 20 Years of Molecular Systematics in Legumes: Reconciling DNA and Morphology into a New Classification System.

Rooting the legumes with full plastid proteome sequence data

Erik Koenen[1], Royce Steeves[2], Anne Bruneau[2], Jan Wieringa[3], Freek Bakker[4], Colin Hughes[1]
[1]Institute of Systematic Botany, University of Zurich, Switzerland
[2]Institut de recherche en biologie végétale, Université de Montreal, Canada
[3]Naturalis Biodiversity Center, Leiden, the Netherlands
[4]Biosystematics department, University of Wageningen, the Netherlands

Many prominent clades (e.g. placental mammals, Angiospermae) have proven difficult to root, particularly when the stem lineage is long, and the legume family (Leguminosae or Fabaceae) is no exception. We believe the problem lies in the difficulty of reconstructing ancestral sequence evolution along the long stem lineage. Also for some of the longer stemmed ingroup taxa, failure to accurately reconstruct ancestral sequences is causing Long Branch Attraction (LBA) artifacts which add to the problem of inferring the deepest relationships in the legumes and Fabales. We attempt to resolve these issues by analyzing a much larger sequence dataset than used previously, using all protein-coding genes from the chloroplast genome. We align and analyze the translated amino acid sequences with heterogeneous mixture models that describe protein evolution more realistically than conventional models used at the DNA level. When employing different outgroups and different ingroup taxon sampling, the root is placed in different positions with high support. Using an experimental approach to phylogenetics, we test for LBA artifacts by removing fast evolving sites and taxa. We provide a new hypothesis of relationships among the Fabales families and the major lineages of the legumes, but one that needs further testing with alternative data and methods.

53rd Annual Meeting of the Association for Tropical Biology and Conservation 2016, Montpellier, France
Oral presentation of 15 minutes
Symposium: Success of tropical legumes and traits that contribute to their dominance

Temporal diversity dynamics of mimosoid legumes, a key ecological component of global tropical biomes

Erik Koenen[1], Kyle Dexter[2], Jens Ringelberg[1], Elvia de Souza[3], Priscilla de Almeida[4], João Iganci[5], Pétala Ribeiro[4], Marcelo Simon[6], Vanessa Terra[7], Donovan Bailey[8], James Boatwright[9], Gillian Brown[10], Javier Aju[11], Michelle van der Bank[12], Joe Miller[13], Luciano de Queiroz[4], Toby Pennington[14], Colin Hughes[1]

[1] University of Zurich - Department : Institute of Systematic Botany, 8008 Zurich - Switzerland,
[2] University of Edinburgh - Department : School of Geosciences, EH9 3FE Edinburgh - United Kingdom,
[3] Universidade do Estado da Bahia - Department : Departamento de Educa, 48608-240 Paulo Afonso - Brazil,
[4] Universidade Estatal Feira de Santana - Department : Departamenta de Ciências Biológica, 44036-900 Feira de Santana - Brazil,
[5] Universidade Federal do Rio Grande do Sul - Department : Instituto de Biociências, Departamento de Botânica, 91501-970 Porto Alegre - Brazil,
[6] Empresa Brasileira de Pesquisa Agropecuária Embrapa - Department : Parque Estação Biológica, 70770-901 Brasilia - Brazil,
[7] Universidade Federal de Uberlândia - Department : Instituto de Ciências Agrárias, 38400-902
Uberlândia - Brazil,
[8] New Mexico State University - Department : Department of Biology, 88003 Las Cruces NM - USA,
[9] University of the Western Cape - Department : Department of Biodiversity and Conservation Biology, 7535 Bellville - Republic of South Africa,
[10] Queensland Herbarium - Department : Department of Science, Information Technology and Innovation, 4066 Brisbane - Australia,
[11] University of Melbourne - Department : School of Biosciences, 3010 Victoria - Australia,
[12] University of Johannesburg - Department : African Centre for DNA Barcoding, 2006 Auckland Park - Republic of South Africa,
[13] National Science Foundation - Department : Division of Environmental Biology, 22230 Arlington VA - USA,
[14] Royal Botanic Gardens Edinburgh - Department : Department of Science, EH3 5LR Edinburgh - United Kingdom

Background – The mimosoid legumes (Leguminosae-Mimosoideae) are a pantropically distributed clade of c. 3300 species of large rainforest trees and lianas, savanna and seasonally dry forest trees and shrubs, and creeping and geoxylic fire-adapted subshrubs. They occupy a wide ecological amplitude spanning the whole lowland tropics and often constitute abundant or dominant elements in tropical rain forests, seasonally dry forests and savannas. Here we analyse the temporal origins and evolutionary dynamics of extant mimosoid diversity in modern tropical biomes, to gain insights into the origins of tropical biodiversity.

Methods – We construct the largest phylogeny for the group to date by adding densely sampled phylogenetic trees of subclades in a hierarchical fashion onto a well-resolved time-calibrated backbone phylogeny based on Next-Generation Sequencing (NGS) of plastid and nuclear genes. We correct for unsampled diversity by simulation and estimate speciation, extinction and net diversification rates across the phylogeny.

Results – While the clade dates back to at least the Early Eocene, most of the extant diversity is derived from later episodes of diversification from the Early Miocene onwards. Exceptionally high diversity is found in the genus Mimosa (c. 550 spp.) and in a large clade comprising the tribes Ingeae and Acacieae p.p. (c. 2000 spp.), which includes multiple nested radiations.

Discussion – We propose that the temporal diversity dynamics of mimosoid legumes are best explained by punctuated extinction and radiation, which leads to episodic species turnover through time. Our findings are important for a general understanding of the temporal assembly of floras, and indeed whole biotas across – and perhaps beyond – the global tropics.

19th International Botanical Congress 2017, Shenzhen, People's Republic of China
Oral presentation of 20 minutes
Symposium: Phylogenomics and Evolution of Legumes

Solving Difficult Phylogenetic Problems in Leguminosae Using Multiple Genome-scale Sequence Data Sets

Erik Koenen[1,4], Jens Ringelberg[1], Catherine Kidner[2], James Nicholls[2], Toby Pennington[2], Jan Wieringa[3], Freek Bakker[4], Royce Steeves[5], Dario Ojeda Alayon[6], Jérémy Migliore[6], Olivier Hardy[6], Luciano de Queiroz[7], Gillian Brown[8], Gwylim Lewis[9], Anne Bruneau[10] and Colin Hughes[1]

[1]Department of Systematic & Evolutionary Botany, University of Zurich, Switzerland
[2]Royal Botanic Gardens Edinburgh, U.K.
[3]Naturalis Biodiversity Center, Leiden, the Netherlands
[4]Biosystematics department, University of Wageningen, the Netherlands
[5]Fisheries and Oceans Canada, Ottawa, Canada
[6]Université Libre de Bruxelles, Brussels, Belgium
[7]Universidade Federal de Feira de Santana, Bahia, Brazil
[8]The University of Melbourne Herbarium, Melbourne, Australia
[9]Royal Botanic Gardens, Kew, Richmond, U.K.
[10]Institut de recherche en biologie végétale, Université de Montreal, Canada

Several parts of the legume phylogeny suffer from lack of resolution. Here we investigate whether genome-scale data can resolve these problems, with a focus on three specific parts of the legume phylogeny: (1) the root of the Leguminosae and relationships among the six subfamilies, (2) relationships among the major lineages in Caesalpinioideae and (3) the large polytomy formed by the clade that includes the c. 37 genera in the non-monophyletic tribes Ingeae and Acacieae p.p. and c. 1,850 of the c. 3,300 species of mimosoid legumes (Caesalpinioideae-*Mimosoida*). We harness the power of large datasets including fully sequenced genomes, transcriptomes, complete chloroplast exomes/proteomes and hybrid capture sequence data for 967 low copy nuclear genes for the Caesalpinioideae. We explore different phylogenetic reconstruction methods using heterogeneous models of molecular evolution on both DNA and protein sequence data and gene-tree species-tree reconciliation methods. Our results show that the root of the legumes remains problematic to resolve, even with large volumes of data. After dissecting the phylogenetic signal among hundreds of nuclear genes, we find that significant numbers of genes support the different possible rootings. We suggest this might be caused by incomplete lineage sorting, rapid diversification of the early legume lineages, extinction of early legume diversity, whole genome duplication,

or most likely, a combination of these factors. Within the Caesalpinioideae, we can resolve the most probable branching order of the major lineages in the subfamily. Previously, lack of resolution across this part of the tree has hindered progress towards a phylogenetic classification of the legume family, even though it was clear that former subfamily Mimosoideae needed to be included in a recircumscribed Caesalpinioideae to avoid having to recognize a large number of additional, small subfamilies. Our results provide the basis for a new tribal classification, and recognition of additional clades with evolutionary significance (e.g. a petiolar nectary clade and an aggregated pollen clade). Within the large clade combining former tribes Ingeae and Acacieae p.p., large scale plastid and nuclear gene data resolve several major subclades for the first time, supporting some previously hypothesized informal generic alliances based on morphology while rejecting others. Furthermore, substantial geographic and ecological structure is apparent across this large clade, but this requires further investigation to understand the relative contributions of dispersal limitation and niche conservatism in shaping the mimosoid phylogeny. Future research will focus on mimosoids, and aims to infer a robust backbone phylogeny for the clade with complete sampling of genera and dense sampling of larger genera using the same set of genes via further hybrid capture work. Our results have implications for revising the tribal classification of the legume family and our study has wider significance in terms of methods for inferring phylogenies of large (plant) clades combining different types of (genome-scale) molecular data.

7th International Legume Conference 2018, Sendai, Japan
Oral presentation of 40 minutes
Plenary talk

Phylogenomic complexity and legume evolution in deep time

Erik Koenen
Institute of Systematic and Evolutionary Botany, University of Zurich

Although the study of legume evolution and phylogeny has progressed tremendously in recent decades, there is still uncertainty concerning resolution of the deepest divergences and the age of the family. Furthermore, there is phylogenetic uncertainty within several subfamilies, perhaps nowhere more so than within mimosoids, where relationships among the c. 40 genera of the Ingeae/Acacieae p.p. clade are almost entirely unresolved in phylogenies based on one or a few markers. Improved estimates of phylogeny and divergence times are needed to investigate how key legume traits evolved and how the enormous contemporary species diversity arose. Analyzing genome-scale molecular sequence data sets, as pursued here, is a promising means of solving difficult phylogenetic problems. For the deepest divergences within the family, results indicate strong disagreement among gene trees about the relationships between subfamilies, suggesting a complex origin of the legumes involving incomplete lineage sorting and/or hybridization linked to ancient whole genome duplication. Hybrid capture data for mimosoids show that the Ingeae/Acacieae p.p. clade is resolved into several subclades but persistent lack of phylogenetic signal among the large majority of genes leaves relationships among these subclades still unresolved. This suggests that initial diversification of this clade occurred nearly instantaneously, such that the backbone of the clade should perhaps be seen as a hard polytomy. Time calibration analyses indicate that legumes likely originated in the Maastrichtian (Late Cretaceous) or possibly the early Paleocene and suggest that early legume evolution is closely associated with polyploidy and the Cretaceous-Paleogene (K-Pg) boundary. During the Cenozoic, legumes appear to have experienced multiple pulses of fast diversification, in line with evidence from the fossil record. Simulations using time-varying birth-death models under different extinction scenarios suggest significant episodic species turnover through time.

Appendix III Supplementary information for Chapter I

Table S1. Accession information for taxa included in the chloroplast alignment.

Table S2. Accession information for taxa included in the nuclear genomic and transcriptomic data set.

Table S3. Counts of bipartitions representing nodes A-H (Fig. 3) and conflicting bipartitions representing other subfamily relationships among 3,473 gene trees.

Figure S1. ML topology as inferred by RAxML from amino acid alignment of chloroplast genes under the LG4X model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

Figure S2. Bayesian majority-rule consensus tree inferred with Phylobayes from amino acid alignment of chloroplast genes under the CATGTR model. Numbers on nodes indicate posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

Figure S3. ML topology as inferred by RAxML from nucleotide alignment of chloroplast genes under the GTR + G model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

Figure S4. Bayesian majority-rule consensus tree inferred with Phylobayes from nucleotide alignment of chloroplast genes under the CATGTR model. Numbers on nodes indicate the posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

Figure S5. ML topology as inferred by RAxML from a concatenated alignment of 1,103 nuclear genes, under the LG4X model. Numbers on nodes indicate Internode Certainty All (ICA) values, as estimated from gene trees of the same 1,103 genes.

Figure S6. Bayesian gene jackknifing majority-rule consensus tree inferred with Phylobayes from a concatenated alignment of 1,103 nuclear genes. Numbers on nodes indicate posterior probabilities (pp), averaged over 500 posterior trees each, for 25 replicates (12,500 posterior trees in total).

Figure S7. Phylogeny estimated under the multi-species coalescent with ASTRAL. Support values indicated represent local posterior probability (blue rectangles) and quartet support (yellow rectangles).

Table S1. Accession information for the taxa included in the chloroplast alignment.

Taxon	Herbarium voucher	Genbank accession number	Comments
Abarema jupunba	M.F. Simon 1600 (CEN)	XXXXXXXXXX	**Newly sequenced**
Acacia koa		see Table S1.	**Transcriptome**
Acacia ligulata		LN555649.2	
Acrocarpus fraxinifolius			**Transcriptome, available at https://ics.hutton.ac.uk/tropiTree/**
Adenanthera pavonina	Ambriansyah & Arifin AA295 (K)	XXXXXXXXXX	**Newly sequenced**
Afzelia africana	S.L.A. Donkpegan 27 (BRLU)	KX673213	
Afzelia bipindensis	S.L.A. Donkpegan 626 (BRLU)	XXXXXXXXXX	**Newly sequenced**
Ajuga reptans		KF709391	
Albizia adianthifolia	J.J. Wieringa 6278 (WAG)	XXXXXXXXXX	**Newly sequenced**
Albizia julibrissin	E. Koenen 601 (Z)	XXXXXXXXXX	**Newly sequenced; Transcriptome**
Anthonotha fragrans		see Table S1.	**Newly sequenced; Transcriptome**
Apios americana		KF856618	
Arabidopsis thaliana		AP000423	
Arachis hypogaea		KJ468094	
Arachis ipaensis		GBIW00000000	**Transcriptome**
Aralia undulata	R. Li 551 (KUN)	KC456163	
Archidendron lucidum	Wang & Lin 2534 (L)	XXXXXXXXXX	**Newly sequenced**
Astragalus membranaceus			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Astragalus propinquus			**Transcriptome, OneKP: MYMP, available at**

			http://www.onekp.com/public_data.html
Azadirachta indica		KF986530	
Bauhinia tomentosa		see Table S1.	**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Bituminaria bituminosa		see Table S1.	**Transcriptome, OneKP: TVSH, available at http://www.onekp.com/public_data.html**
Bulnesia arborea	M.J. Moore 334 (FLAS)	EU002159, EU002172, EU002205, EU002275, EU002299, EU002388, EU002478, GQ998005-GQ998073, HQ664597	
Buxus microphylla		EF380351	
Calliandra hygrophila	L.P. Queiroz 15542 (HUEFS)	XXXXXXXXXX	**Newly sequenced**
Carica papaya		EU431223	
Ceratonia siliqua		KJ468096	
Cercis canadensis		KF856619	
Chamaecrista fasciculata		see Table S1.	**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Chidlowia sanguinea	J.J. Wieringa 4338 (WAG)	XXXXXXXXXX	**Newly sequenced**
Chrysobalanus icaco		KJ414480	
Cicer arietinum		EU835853	
Citrus sinensis		DQ864733	
Cladrastis lutea			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Codariocalyx motorius			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Coffea arabica		EF044213	

Cojoba arborea	M.F. Simon 1545 (CEN)	XXXXXXXXXX	**Newly sequenced**
Colvillea racemosa	Kew living collection 1993-224 (K)	XXXXXXXXXX	**Newly sequenced**
Copaifera officinalis			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Cucumis sativus		AJ970307	
Daucus carota		DQ898156	
Desmanthus illlinoiensis			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Dialium guineense	T. van Andel 4184 (WAG)	XXXXXXXXXX	**Newly sequenced**
Dichrostachys cinerea	O. Maurin 256 (JRAU)	XXXXXXXXXX	**Newly sequenced**
Dimorphandra macrostachya	J.R. Igançi 877 (RB)	XXXXXXXXXX	**Newly sequenced**
Diptychandra aurantiaca	J.R.I. Wood 26513 (K)	XXXXXXXXXX	**Newly sequenced**
Distemonanthus benthamianus	G. Dauby 728 (BRLU)	XXXXXXXXXX	**Newly sequenced**
Duparquetia orchidacea	J.J. Wieringa 7805 (L)	XXXXXXXXXX	**Newly sequenced**
Entada abyssinica	MSB 0133199 (K)	XXXXXXXXXX	**Newly sequenced; Transcriptome**
Entada rheedei	E. Koenen 496 (Z)	XXXXXXXXXX	**Newly sequenced**
Erythrophleum ivorense	J.J. Wieringa 5487 (WAG)	XXXXXXXXXX	**Newly sequenced**
Erythrostemon gilliesii	R. Steeves 852 (MT)	XXXXXXXXXX	**Newly sequenced**
Eucalyptus grandis		HM347959	
Euonymus americanus	W. Judd 8071 (FLAS)	EU002160, EU002170, EU002193, EU002277, EU002321, EU002409, EU002500, GQ998147-GQ998219, HQ664608	

Fagopyrum esculentum		EU254477	
Faidherbia albida	O. Maurin 3495 (JRAU)	XXXXXXXXXX	**Newly sequenced**
Garcinia mangostana		HQ331601, HQ331906, HQ332057, HQ848709, JX661816, JX661859, JX661902, JX661944, JX661980, JX662020, JX662065, JX662109, JX662151, JX662196, JX662237, JX662279, JX662320, JX662359, JX662399, JX662434, JX662467, JX662502, JX662543, JX662580, JX662622, JX662666, JX662710, JX662752, JX662799, JX662841, JX662880, JX662914, JX662955, JX662996, JX663032, JX663071, JX663104, JX663149, JX663196, JX663237, JX663280, JX663322, JX663365, JX663410, JX663583, JX663630, JX663677, JX663721, JX663763, JX663804, JX663841, JX663874, JX663915, JX663962, JX664006, JX664049, JX664091, JX664127, JX664165, JX664209, JX664252, JX664297, JX664341, JX664385,	

		JX664458, JX664495, JX664535, JX664580, JX664623, JX664659, JX664694, JX664726, JX664771, JX664812, JX664852, JX664895, JX664939, JX665004, KF783277, U92876, U92877, U92878	
Gleditsia sinensis			**Transcriptome, OneKP: VHZV, available at http://www.onekp.com/public_data.html**
Gleditsia triacanthos			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Gliricidia sepium			**Transcriptome, available at https://ics.hutton.ac.uk/tropiTree/**
Glycine canescens		KC893635	
Glycine max		DQ317523	
Glycyrrhiza glabra		KF201590	
Glycyrrhiza lepidota			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Gompholobium polymorphum			**Transcriptome, OneKP: VLNB, available at http://www.onekp.com/public_data.html**
Gossypium hirsutum		DQ345959	
Guibourtia ehie	F. Tosso 272 (BRLU)	XXXXXXXXXX	**Newly sequenced**
Guibourtia tessmannii		XXXXXXXXXX	**Newly sequenced**
Guilfoylia monostylis	P.I. Forster 28103 (Z)	XXXXXXXXXX	**Newly sequenced**
Gymnocladus dioicus			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Haematoxylum brasiletto		KJ468097	
Helianthus annuus		DQ383815	

Hymenostegia brachyura	Zenker 4481 (WAG)	XXXXXXXXXX	**Newly sequenced**
Hymenostegia felicis	Jacques-Félix 5129 (WAG)	XXXXXXXXXX	**Newly sequenced**
Indigofera tinctoria		KJ468098	
Inga leiocalycina	T.D. Pennington 13822 (K)	KT428296	
Inga spectabilis	T.D. Pennington 15061 (K)	XXXXXXXXXX	**Newly sequenced**
Intsia bijuga		KX673214	
Lathyrus graminifolius		KJ806193	
Lathyrus sativus		HM029371	
Lens culinaris		KF186232	
Leucaena trichandra		KT428297	
Libidibia coriaria		KJ468095	
Lotus japonicus		AP002983	
Lupinus luteus		KC695666	
Lupinus polyphyllus			**Transcriptome, OneKP: CMFF, available at http://www.onekp.com/public_data.html**
Macadamia integrifolia		KF862711	
Manihot esculenta		EU117376	
Medicago hybrida		KJ850240	
Medicago truncatula		AC093544	
Microlobius foetidus	C.E. Hughes 1219 (FHO)	XXXXXXXXXX	**Newly sequenced; Transcriptome**
Millettia pinnata		JN673818	
Mimosa tenuiflora	L.P. Queiroz 15498 (HUEFS)	XXXXXXXXXX	**Newly sequenced**
Morus indica		DQ226511	
Nelumbo nucifera		JQ336993	
Nerium oleander	W. Judd 8076 (FLAS)	KJ953907	

Newtonia hildebrandtii	O. Maurin 2457 (JRAU)	XXXXXXXXXX	**Newly sequenced**
Oenothera biennis		EU262889	
Olea europaea		GU228899	
Oxalis latifolia	M.J. Moore 316 (FLAS)	EU002165, EU002186, EU002248, EU002282, EU002350, EU002438, EU002528, GQ998511-GQ998580, HQ664602, KF783277, U92876, U92877, U92878	
Pachyelasma tessmannii	J.J. Wieringa 5229 (WAG)	XXXXXXXXXX	**Newly sequenced**
Pachyrhizus erosus		KJ468100	
Paeonia obovata		KJ206533	
Parkia panurensis	J.R. Igançi 842 (RB)	XXXXXXXXXX	**Newly sequenced**
Pelargonium alternans		KF240617	
Peltophorum africanum	Koenen 601 (Z)	XXXXXXXXXX	**Newly sequenced**
Pentaclethra macrophylla	Galeuchet & Balthazar 10 (Z)	XXXXXXXXXX	**Newly sequenced**
Phaseolus vulgaris		DQ886273	
Piptadeniastrum africanum	E. Koenen 152 (WAG)	XXXXXXXXXX	**Newly sequenced**
Pisum sativum		HM029370	
Pithecellobium dulce	B. Marazzi 309 (?)	XXXXXXXXXX	**Newly sequenced**
Poeppigia procera	Hernández 558 (Z)	XXXXXXXXXX	**Newly sequenced**
Polygala lutea		EF489041	
Populus trichocarpa		KF753634	
Primula poissonii		KF753634	

Prioria balsamifera		XXXXXXXXXX	**Newly sequenced; Transcriptome**
Prosopis alba		see Table S1.	**Transcriptome**
Prosopis glandulosa		KJ468101	
Prunus persica		HQ336405	
Quercus rubra		JX970937	
Quillaja saponaria			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Ranunculus macranthus		DQ359689	
Robinia pseudoacacia		KJ468102	
Samanea saman	C.E. Hughes 421 (FHO)	XXXXXXXXXX	**Newly sequenced**
Sapindus mukorossi		KM454982	
Saraca indica	Kew living collection 2011-1421 (K)	XXXXXXXXXX	**Newly sequenced**
Schotia brachypetala	R. Steeves 846 (MT)	XXXXXXXXXX	**Newly sequenced**
Sedum sarmentosum		JX427551	
Senegalia ataxacantha	C. Jongkind 10603 (WAG)	XXXXXXXXXX	**Newly sequenced**
Senna hebecarpa			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Senna siamea			**Transcriptome, available at https://ics.hutton.ac.uk/tropiTree/**
Sesbania macrantha			**Transcriptome, available at https://ics.hutton.ac.uk/tropiTree/**
Sesbania sesban			**Transcriptome, available at https://ics.hutton.ac.uk/tropiTree/**
Silene latifolia		JF715055	
Solanum lycopersicum		KP331414	
Styphnolobium japonicum		see Table S1.	**Transcriptome, available at https://www.hindawi.com/journals/bmri/2014/75**

			0961/sup/
Swartzia emarginata	M.P. Morim 576 (RB)	XXXXXXXXXX	**Newly sequenced**
Tachigali odoratissima	M.P. Morim 562 (RB)	XXXXXXXXXX	**Newly sequenced**
Tamarindus indica		KJ468103	
Theobroma cacao		HQ244500	
Tipuana tipu			**Transcriptome, available at https://ics.hutton.ac.uk/tropiTree/**
Trachelium caeruleum		EU090187	
Trifolium aureum		KC894708	
Trifolium repens		KC894706	
Trochodendron aralioides		KC608753	
Vaccinium macrocarpon		JQ757046	
Vachellia tortilis	E. Koenen 603 (Z)	XXXXXXXXXX	**Newly sequenced**
Vicia faba		KF042344	
Vicia sativa		KJ850242	
Vigna radiata		GQ893027	
Vigna unguiculata		JQ755301	
Vitis vinifera		DQ424856	
Wisteria floribunda			**Transcriptome, OneKP: RMWJ, available at http://www.onekp.com/public_data.html**
Xanthocercis zambesiaca			**Transcriptome, available at http://dx.doi.org/10.5061/dryad.ff1tq**
Xanthophyllum eurhynchum	P. Herendeen H.416 (?)	XXXXXXXXXX	**Newly sequenced**
Xylia hoffmannii	E. Koenen 402 (Z)	XXXXXXXXXX	**Newly sequenced**
Zenia insignis	Averyanov et al. 5748 (?)	XXXXXXXXXX	**Newly sequenced**

Table S2. Accession information for the taxa included in the nuclear genomic and transcriptomic data set.

Taxon	Source	Citation
Acacia koa	Genbank BioProject: PRJNA268386	Ishihara et al.
Acrocarpus fraxinifolius	TropiTree: https://ics.hutton.ac.uk/tropiTree/	Russel et al. 2014
Afzelia bella	Genbank BioProject: XXXXXXXXXX	**Newly sequenced**
Albizia julibrissin	Genbank BioProject: XXXXXXXXXX	**Newly sequenced**
Alnus serrulata	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Amaranthus hypochondriacus	Phytozome v11: v1.0	Clouse et al. 2016
Anthonotha fragrans	Genbank BioProject: XXXXXXXXXX	**Newly sequenced**
Apios americana	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Aquilegia coerulea	Phytozome v11: v3.1	Filiault et al. 2018
Arabidopsis thaliana	Phytozome v11: TAIR10	Lamesch et al. 2012
Arachis ipaensis	Peanutbase.org: K30076.a1.M1	Bertioli et al. 2016
Astragalus membranaceus	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Bauhinia tomentosa	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Bituminaria bituminosa	OneKP: TVSH	Wicket et al. 2014
Cajanus cajan	http://gigadb.org/dataset/100028	Varshney et al. 2012
Cannabis sativa	Genbank BioProject: PRJNA74271	van Bakel et al. 2011
Carica papaya	Phytozome v11: ASGPBv4.0	Ming et al. 2008
Castanea mollissima	https://www.hardwoodgenomics.org/Genome-assembly/1962958	*Not available*
Cercis canadensis	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Chamaecrista fasciculata	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Cicer arietinum	http://gigadb.org/dataset/100076	Varshney et al. 2013
Citrus sinensis	Phytozome v11: v1.1	Wu et al. 2014

Cladrastis lutea	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Codariocalyx motorius	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Copaifera officinalis	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Cucumis sativus	Phytozome v11: v1.0	*Not available*
Desmanthus illinoensis	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Elaeocarpus photiniifolia	Genbank BioProject: PRJDA67329	Sugai et al. 2012
Entada abyssinica	Genbank BioProject: XXXXXXXXXX	**Newly sequenced**
Eucalyptus grandis	Phytozome v11: v2.0	Bartholomé et al. 2015
Fragaria vesca	Phytozome v11: v1.1	Shulaev et al. 2011
Gleditsia triacanthos	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Glycine max	Phytozome v11: Wm82.a2.v1	Schmutz et al. 2010
Glycyrrhiza lepidota	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Gossypium raimondii	Phytozome v11: v2.1	Paterson et al. 2012
Gymnocladus dioicus	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Inga spectabilis	https://doi.org/10.5061/dryad.r9c12	Nicholls et al. 2015
Juglans regia	https://www.hardwoodgenomics.org/Genome-assembly/2209485	Martínez-García et al. 2016
Lactuca sativa	Genbank BioProject: PRJNA65477	*Not available*
Lathyrus sativus	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Lens culinaris	Genbank BioProject: PRJNA65667	Kaur et al. 2011
Linum usitatissimum	Phytozome v11: v1.0	Wang et al. 2012
Lotus japonicus	http://www.plantgdb.org/LjGDB/	Sato et al. 2008
Lupinus angustifolius	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Lupinus polyphyllus	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Manihot esculenta	Phytozome v11: v6.1	Bredeson et al. 2016
Medicago truncatula	Phytozome v11: Mt4.0v1	Young et al. 2011

Microlobius foetidus	Genbank BioProject: XXXXXXXXXX	**Newly sequenced**
Mimulus guttatus	Phytozome v11: v2.0	Hellsten et al. 2013
Morus notabilis	Genbank BioProject: PRJNA202089 (assembly version ASM41409v2)	He et al. 2013
Nelumbo nucifera	Genbank BioProject: PRJNA264089 (assembly version 1.1)	Ming et al. 2013
Paeonia lactiflora	Genbank BioProject: PRJNA245064	Zhang et al. 2015
Panax ginseng	Genbank BioProject: PRJNA173906	Li et al. 2013
Papaver somniferum	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Phaseolus vulgaris	Phytozome v11: v1.0	Schmutz et al. 2014
Pisum sativum	Genbank BioProject: PRJNA211622	Duarte et al. 2014
Polygala lutea	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Populus trichocarpa	Phytozome v11: v3.0	Tuskan et al. 2006
Primula veris	https://doi.org/10.5061/dryad.2s200	Nowak et al. 2015
Prioria balsamifera	Genbank BioProject: XXXXXXXXXX	**Newly sequenced**
Prosopis alba	Genbank BioProject: PRJNA218545	Torales et al. 2013
Prunus persica	Phytozome v11: v2.1	International Peach Genome Initiative et al., 2013
Punica granatum	Genbank BioProject: PRJNA231033	Ophir et al. 2014
Quillaja saponaria	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Salix purpurea	Phytozome v11: v1.0	Zhou et al. 2018
Senna hebecarpa	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Solanum tuberosum	Phytozome v11: v3.4	Sharma et al. 2013
Styphnolobium japonicum	https://www.hindawi.com/journals/bmri/2014/750961/sup/	Zhu et al. 2014
Theobroma cacao	Phytozome v11: v1.1	Motamayor et al. 2013
Trifolium pratense	Genbank BioProject: PRJNA219226	Yates et al. 2014
Tripterygium wilfordii	Genbank BioProject: PRJNA218574	*Not available*

Vicia faba	Genbank BioProject: PRJNA81211	Kaur et al. 2012
Vigna radiata	ftp://plantgenomics.snu.ac.kr/mungbean_data/	Kang et al. 2014
Vitis vinifera	Phytozome v11: Genoscope.12X	Jaillon et al. 2007
Xanthocercis zambesiaca	Dryad: http://dx.doi.org/10.5061/dryad.ff1tq	Cannon et al. 2015
Zenia insignis	Genbank BioProject: PRJNA285444	*Not available*

Phytozome is available at https://phytozome.jgi.doe.gov
OneKP data is available at http://www.onekp.com/public_data.html

References

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, et al. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. Nat Genet. 2016 Apr;48(4):438–446.

Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat Biotechnol. 2016 May;34(5):562–570.

Clouse JW, Adhikary D, Page JT, Ramaraj T, Deyholos MK, Udall JA, et al. The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly. Plant Genome. 2016 Mar;9:1.

Duarte J, Rivière N, Baranger A, Aubert G igoire, Burstin J, Cornet L, et al. Transcriptome sequencing for high throughput SNP development and genetic mapping in Pea. BMC Genomics. 2014 Feb;15:126.

Filiault DL, Ballerini ES, Mandáková T, Aköz G, Derieg NJ, Schmutz J, et al. The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. eLife. 2018 Oct;

He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, et al. Draft genome sequence of the mulberry tree Morus notabilis. Nat Commun. 2013;4:2445.

Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. PNAS. 2013 Nov;110(48):19478–19482.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007 Sep;449(7161):463–467.

Kaur S, Cogan NOI, Pembleton LW, Shinozuka M, Savin KW, Materne M, et al. Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics. 2011 May;12:265.

Kaur S, Pembleton LW, Cogan NOI, Savin KW, Leonforte T, Paull J, et al. Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. BMC Genomics. 2012 Mar;13:104.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012 Jan;40(Database):issue.

Li C, Zhu Y, Guo X, Sun C, Luo H, Song J, et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in Panax ginseng C. A. Meyer. BMC Genomics. 2013 Apr;14:245.

Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, et al. The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. Plant J. 2016 Sep;87(5):507–532.
Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, et al. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature. 2008 Apr;452(7190):991–996.
Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, et al. Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biol. 2013 May;14(5):R41.
Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, et al. Phylogenetic Analysis of the Plastid Inverted Repeat for 244 Species: Insights into Deeper-Level Angiosperm Relationships from a Long, Slowly Evolving Sequence Region. International Journal of Plant Sciences. 2011 May;172(4):541–558.
Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. PNAS. 2010 Mar;107(10):4623–4628.
Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013 Jun;14(6):r53.
Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of Eucalyptus grandis. Nature. 2014 Jun;510(7505):356–362.
Nicholls JA, Pennington RT, Koenen EJ, Hughes CE, Hearn J, Bunnefeld L, et al. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae). Front Plant Sci. 2015 Sep;6.
Nowak MD, Russo G, Schlapbach R, Huu CN, Lenhard M, Conti E. The draft genome of Primula veris yields insights into the molecular basis of heterostyly. Genome Biol. 2015 Jan;16:12.
Ophir R, Sherman A, Rubinstein M, Eshed R, Sharabi Schwager M, Harel-Beja R, et al. Single-nucleotide polymorphism markers from de-novo assembly of the pomegranate transcriptome reveal germplasm genetic diversity. PLoS One. 2014 Feb;9(2):e88998.
Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature. 2012 Dec;492(7429):423–427.
Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010 Jan;463(7278):178–183.
Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014 Jun;46(7):707.
Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, Carboni MF, et al. Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 (Bethesda). 2013 Nov;3(11):2031–2047.
Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (Fragaria vesca). Nat Genet. 2011 Feb;43(2):109–116.
Sugai K, Setsuko S, Uchiyama K, Murakami N, Kato H, Yoshimaru H. Development of EST-SSR markers for Elaeocarpus photiniifolia (Elaeocarpaceae), an endemic taxon of the Bonin Islands. Am J Bot. 2012 Feb;99(2):84–87.

Torales SL, Rivarola M iximo, Pomponio M ia F, Gonzalez S, Acuña CV, Fernández P, et al. De novo assembly and characterization of leaf transcriptome for the development of functional molecular markers of the extremophile multipurpose tree species Prosopis alba. BMC Genomics. 2013 Oct;14:705.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006 Sep;313(5793):1596–1604.

van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, et al. The draft genome and transcriptome of Cannabis sativa. Genome Biol. 2011 Oct;12(10):R102.

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, et al. Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol. 2013 Mar;31(3):240–246.

Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, et al. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci USA. 2009 Mar;106(10):3853–3858.

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, et al. The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. Plant J. 2012 Nov;72(3):461–473.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci USA. 2014 Nov;111(45):E4859–E4868.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014 Jul;32(7):656–662.

Xi Z, Ruhfel BR, Schaefer H, Amorim A i M, Sugumaran M, Wurdack KJ, et al. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. PNAS. 2012 Oct;109(43):17519–17524.

Yates SA, Swain MT, Hegarty MJ, Chernukin I, Lowe M, Allison GG, et al. De novo assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. BMC Genomics. 2014 Jun;15:453.

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, et al. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. 2011 Nov;480(7378):520–524.

Zhang J, Wu Y, Li D, Wang G, Li X, Xia Y. Transcriptomic analysis of the underground renewal buds during dormancy transition and release in “Hangbaishao” peony (Paeonia lactiflora). PLoS One. 2015 Mar;10(3):e0119118.

Zhou R, Macaya-Sanz D, Rodgers-Melnick E, Carlson CH, Gouker FE, Evans LM, et al. Characterization of a large sex determination region in Salix purpurea L. (Salicaceae). Mol Genet Genomics. 2018 Jul;1–16.

Zhu L, Zhang Y, Guo W, Xu X-J, Wang Q. De Novo Assembly and Characterization of Sophora japonica Transcriptome Using RNA-seq. Biomed Res Int. 2014 Jan;2014.

Table S3. Counts and percentages of bipartitions representing nodes A-H and conflicting bipartitions representing other subfamily relationships among 3,473 gene trees.

Clade	ML		>50% bootstrap support		>80% bootstrap support	
	count	percentage	count	percentage	count	percentage
bipartitions of best supported topology						
Leguminosae (node A)	2669	76.85%	2254	64.90%	1660	47.80%
Cerc + Detar (node B)	744	21.42%	325	9.36%	48	1.38%
Cercidoideae (node C)	1815	52.26%	1705	49.09%	1394	40.14%
Detarioideae (node D)	3041	87.56%	2918	84.02%	2585	74.43%
Pap + Caes + Dial (node E)	794	22.86%	360	10.37%	91	2.62%
Pap + Caes (node F)	599	17.25%	231	6.65%	42	1.21%
Caesalpinioideae (node G)	2114	60.87%	1712	49.29%	1151	33.14%
Papilionoideae (node H)	2456	70.72%	1957	56.35%	1248	35.93%
conflicting bipartitions						
Pap + Caes + Dial + Detar	625	18.00%	258	7.43%	34	0.98%
Pap + Caes + Dial + Cerc	546	15.72%	194	5.59%	20	0.58%
Pap + Dial	446	12.84%	133	3.83%	20	0.58%
Caes + Dial	448	12.90%	132	3.80%	16	0.46%
Dial + Cerc	295	8.49%	93	2.68%	7	0.20%
Dial + Detar	307	8.84%	96	2.76%	4	0.12%
Caes + Dial + Cerc + Detar	247	7.11%	47	1.35%	4	0.12%
Pap + Dial + Cerc + Detar	196	5.64%	29	0.84%	4	0.12%
Caes + Detar	200	5.76%	44	1.27%	3	0.09%
Pap + Caes + Cerc + Detar	234	6.74%	46	1.32%	2	0.06%
Pap + Detar	189	5.44%	41	1.18%	2	0.06%
Caes + Cerc	163	4.69%	37	1.07%	2	0.06%
Pap + Cerc	173	4.98%	30	0.86%	2	0.06%
Pap + Caes + Detar	153	4.41%	27	0.78%	1	0.03%
Dial + Cerc + Detar	202	5.82%	21	0.60%	1	0.03%
Pap + Dial + Detar	122	3.51%	12	0.35%	1	0.03%
Caes + Dial + Cerc	121	3.48%	16	0.46%	0	0.00%
Pap + Caes + Cerc	132	3.80%	15	0.43%	0	0.00%
Caes + Cerc + Detar	127	3.66%	14	0.40%	0	0.00%
Pap + Dial + Cerc	110	3.17%	12	0.35%	0	0.00%
Caes + Dial + Detar	134	3.86%	11	0.32%	0	0.00%
Pap + Cerc + Detar	115	3.31%	9	0.26%	0	0.00%

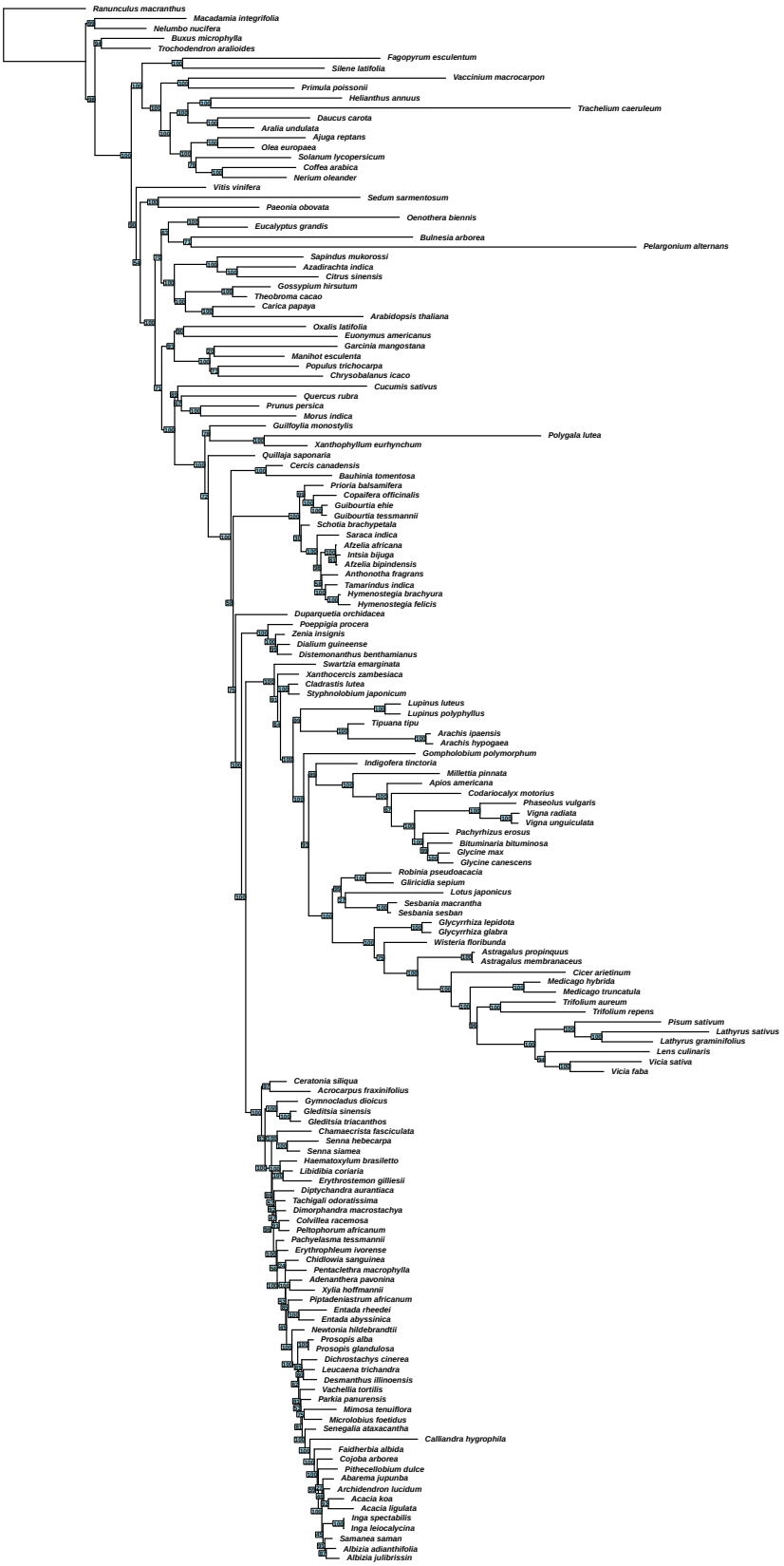


Figure S1. ML topology as inferred by RAxML from amino acid alignment of chloroplast genes under the LG4X model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

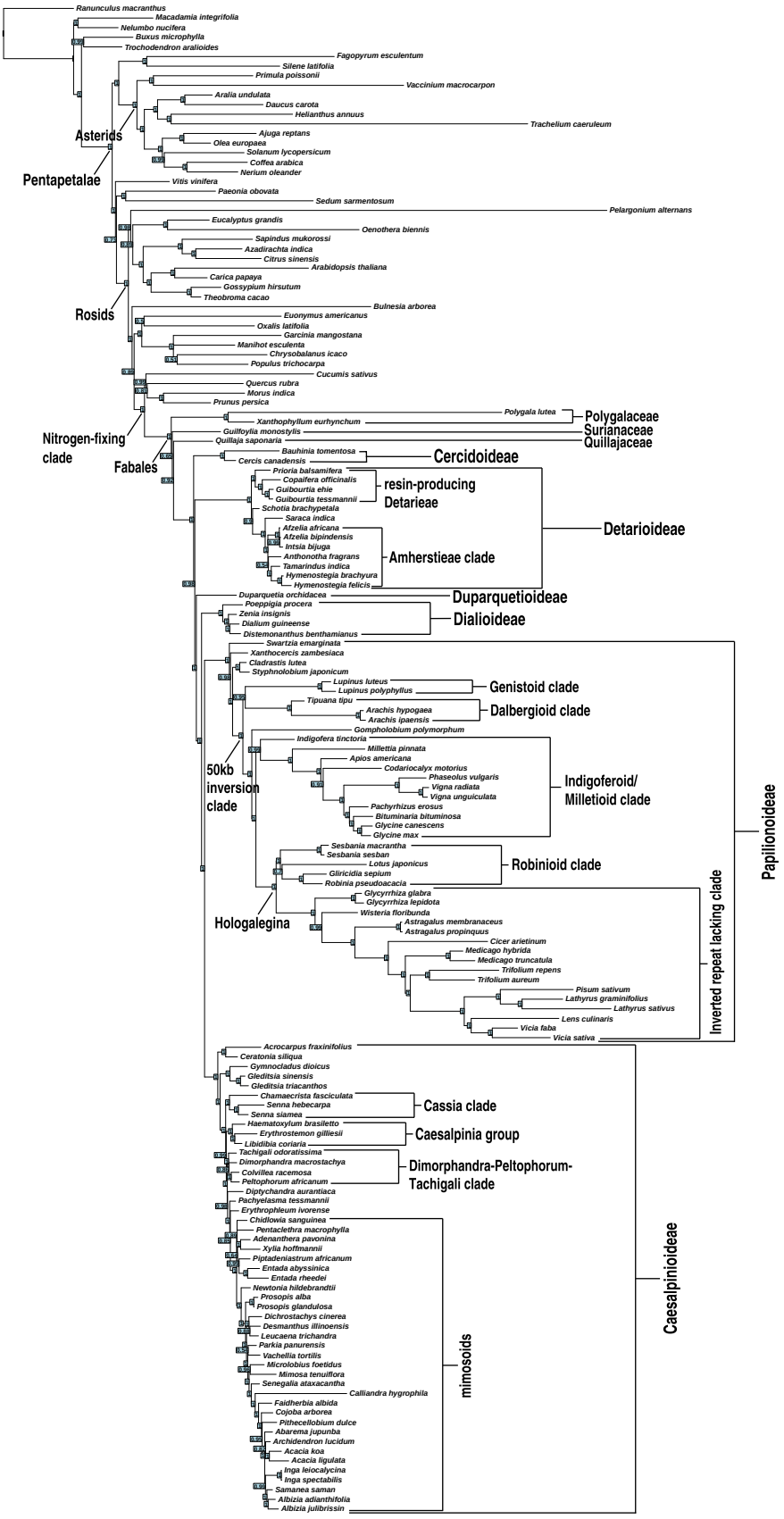


Figure S2. Bayesian majority-rule consensus tree inferred with Phylobayes from amino acid alignment of chloroplast genes under the CATGTR model. Numbers on nodes indicate posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

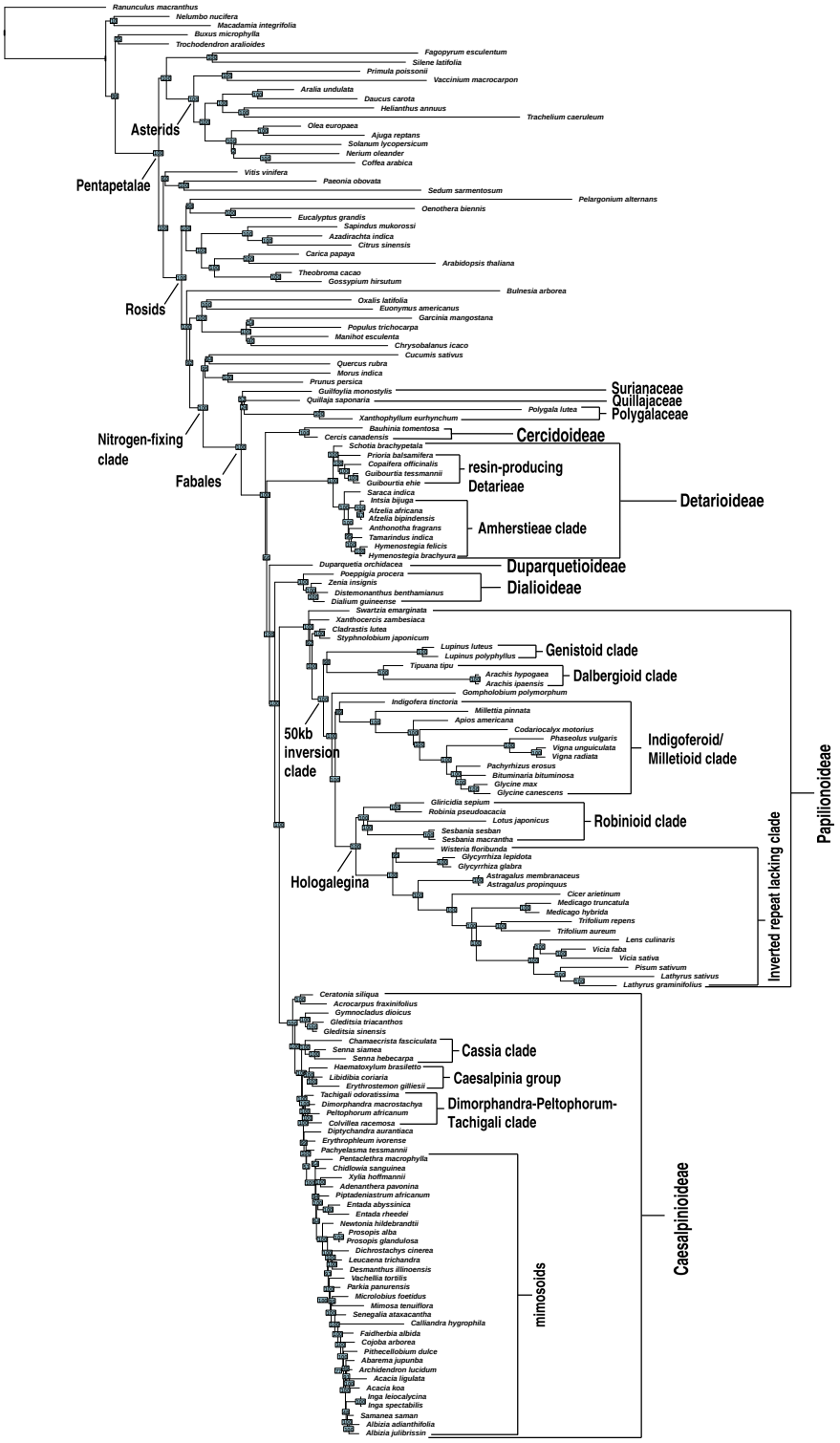


Figure S3. ML topology as inferred by RAxML from nucleotide alignment of chloroplast genes under the GTR + G model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

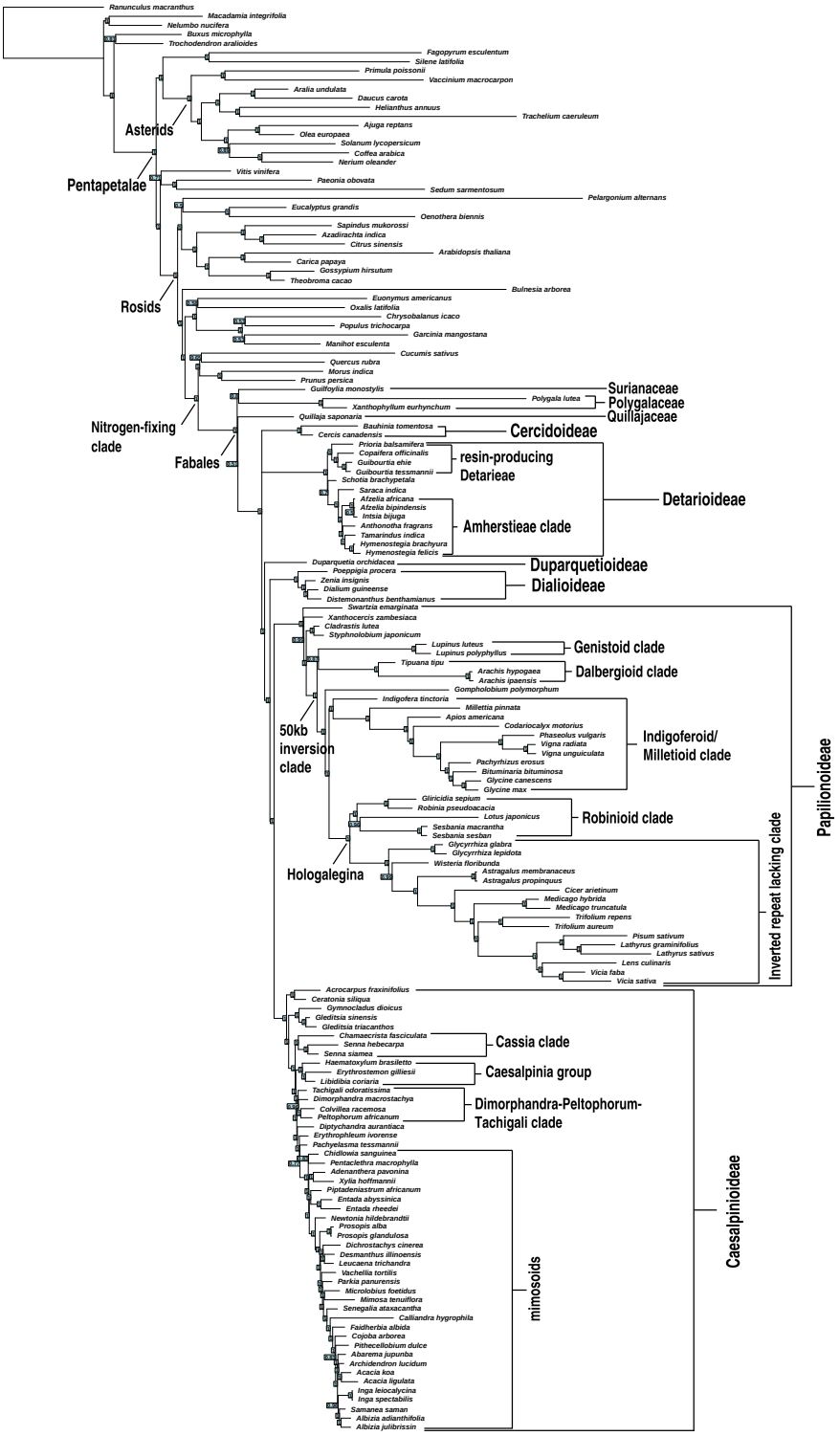


Figure S4. Bayesian majority-rule consensus tree inferred with Phylobayes from nucleotide alignment of chloroplast genes under the CATGTR model. Numbers on nodes indicate the posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

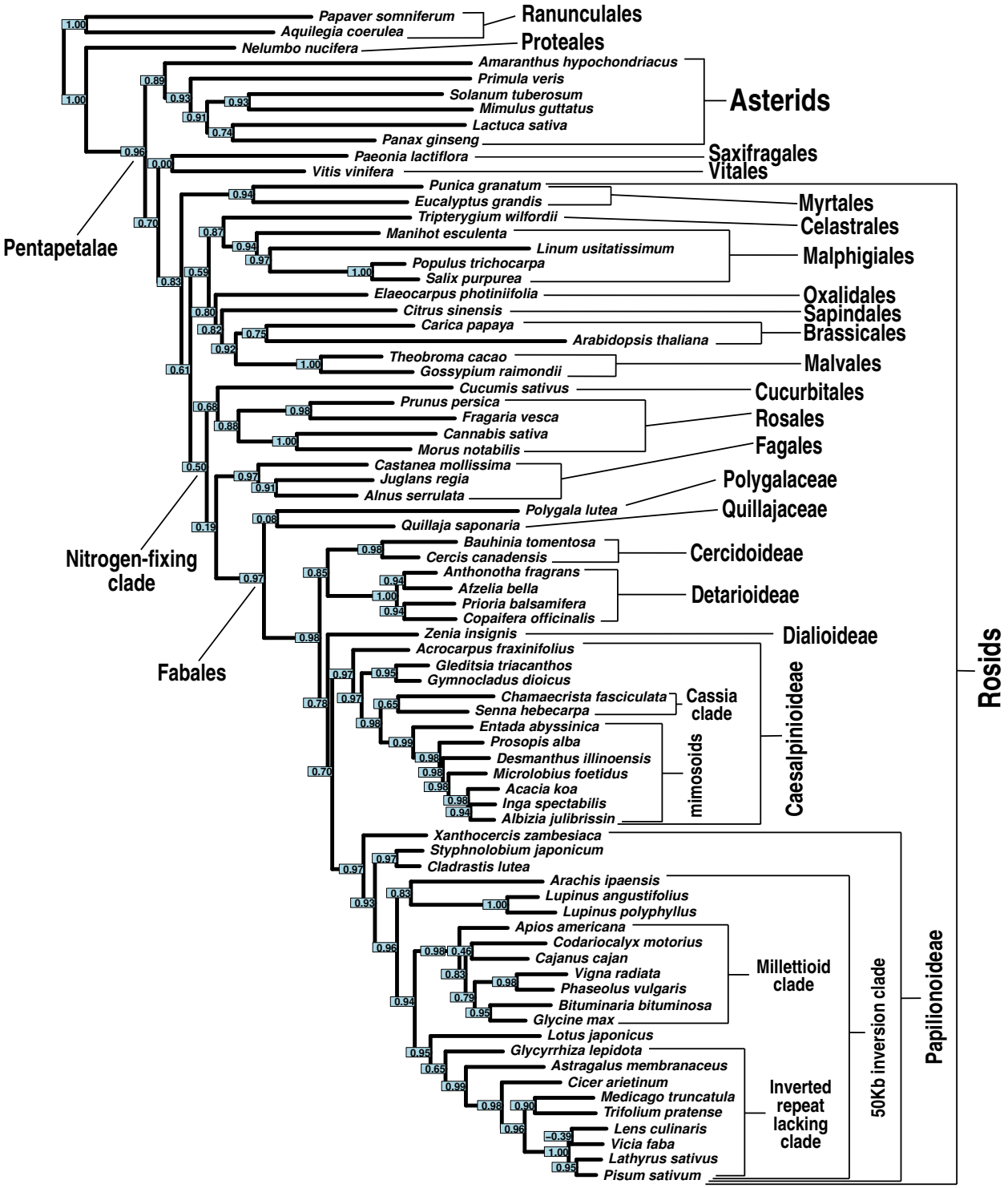


Figure S5. ML topology as inferred by RAxML from a concatenated alignment of 1,103 nuclear genes, under the LG4X model. Numbers on nodes indicate Internode Certainty All (ICA) values, as estimated from gene trees of the same 1,103 genes.

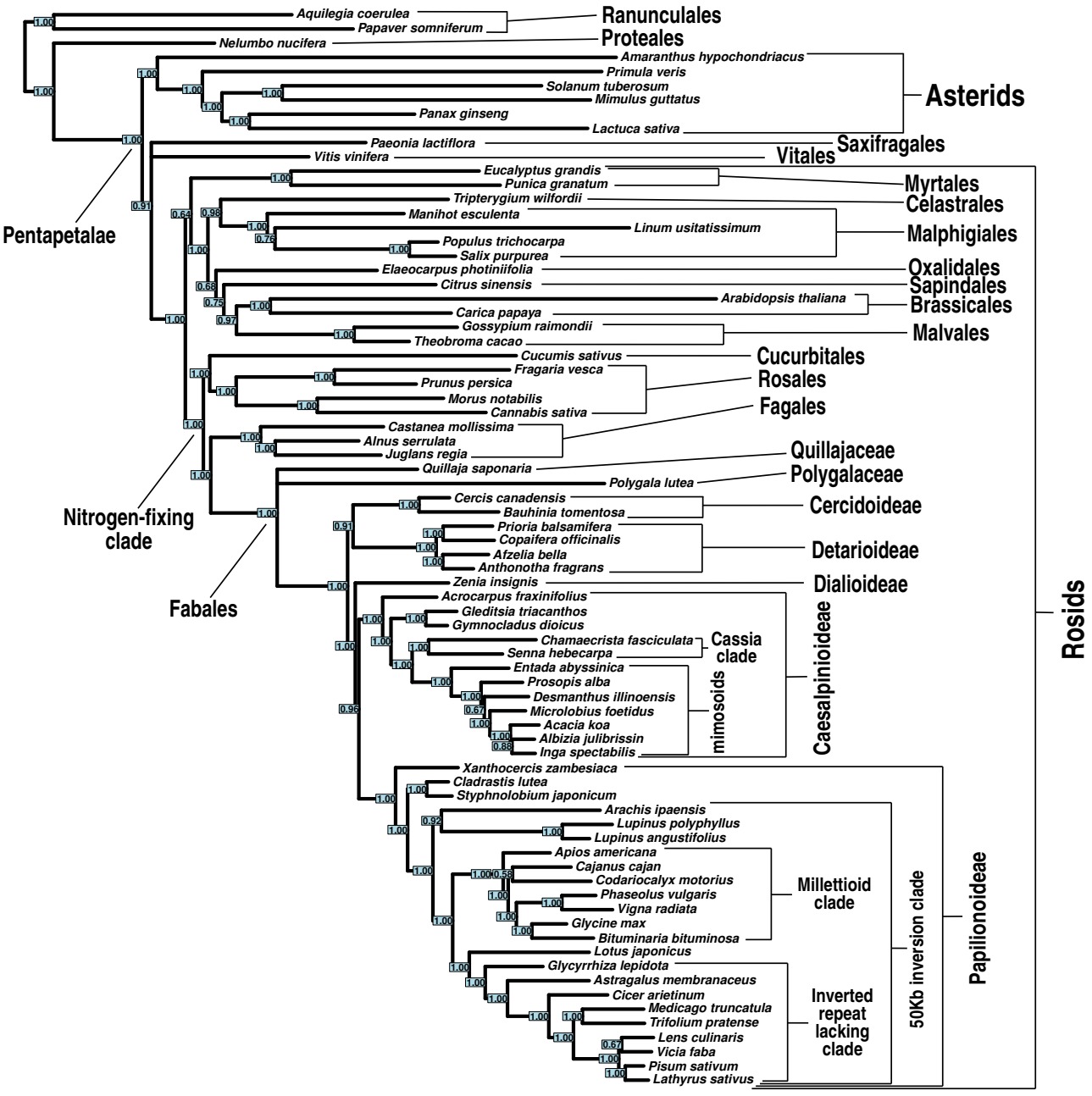


Figure S6. Bayesian gene jackknifing majority-rule consensus tree inferred with Phylobayes from a concatenated alignment of 1,103 nuclear genes. Numbers on nodes indicate posterior probabilities (pp), averaged over 500 posterior trees each, for 25 replicates (12,500 posterior trees in total).

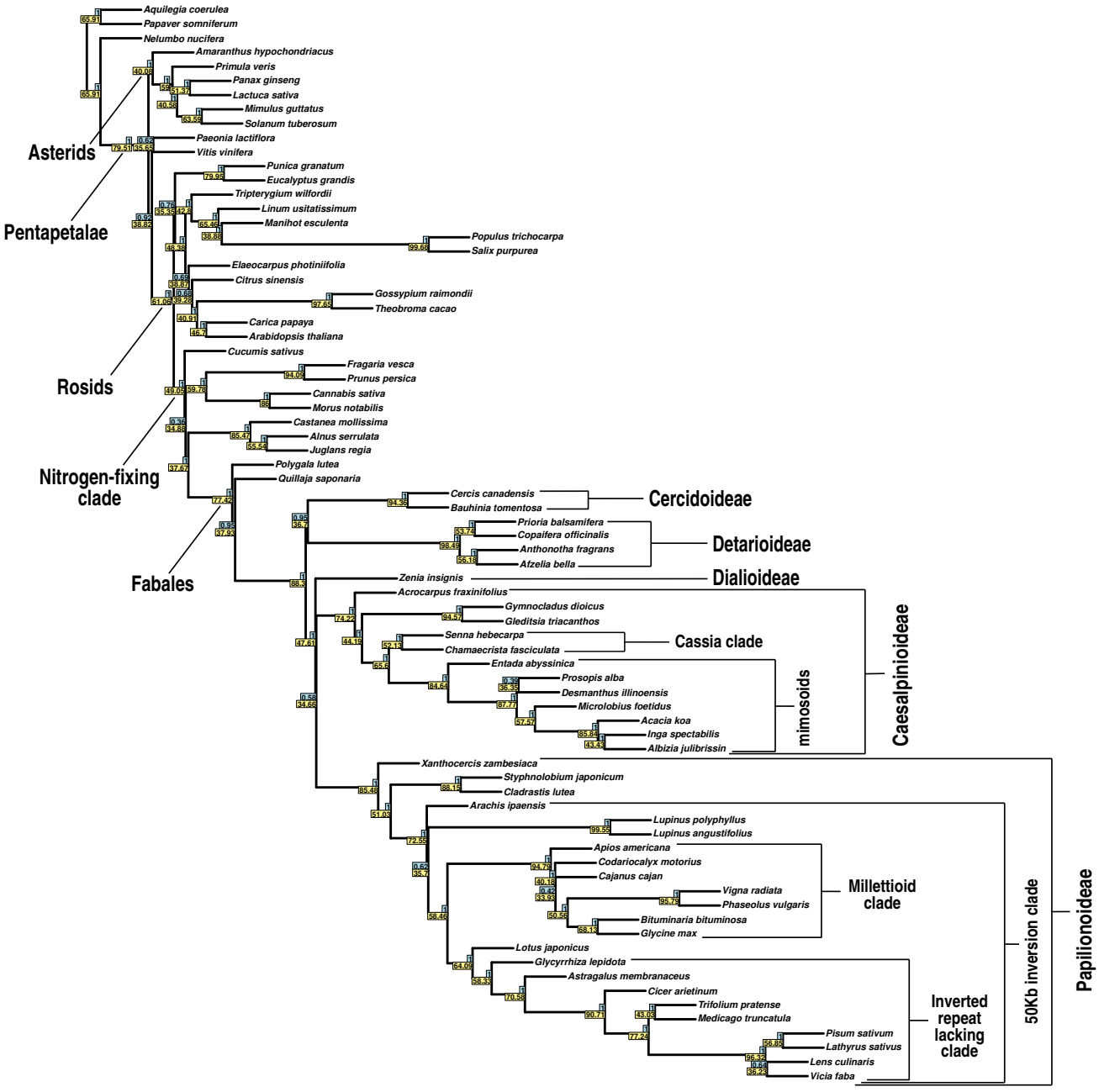


Figure S7. Phylogeny estimated under the multi-species coalescent with ASTRAL. Support values indicated represent local posterior probability (blue rectangles) and quartet support (yellow rectangles).

Appendix IV Supplementary information for Chapter II

Table S1. Taxon occupancy per analysis..

Table S2. Age intervals specified for the fossil calibration priors under different alternative priors.

Table S3. Node age estimates and priors (95% HPD intervals) of nodes A-H in the different analyses.

Figure S1. Examples of homolog clusters with gene duplications in legumes that passed the bootstrap filter. Yellow stars behind nodes indicate locations of gene duplications, numbers on nodes indicate bootstrap support. The plotted gene trees are extracted from (a) cluster3675_1rr_1rr, showing a duplication subtending Detarioideae, (b) cluster1032_1rr_1rr, showing a duplication subtending Papilionoideae, (c) cluster1248_1rr_1rr and (d) cluster2941_1rr_1rr, both with a duplication subtending the legume family. Trees for (e) cluster51_7rr_1rr and (f) cluster544_1rr_1rr show evidence of more than one duplication, including one specific to Papilionoideae in the former.

Figure S2. Numbers of gene duplications mapped across the species tree as estimated by Phyparts. The topology used is the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,038 homolog clusters. Numbers above branches (with blue background) and below branches (with yellow background) represent numbers of duplications and numbers of homolog trees with duplications without or with a bootstrap filter of 50%, respectively.

Figure S3. Numbers of gene duplications mapped across the species tree as estimated by Notung. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,324 homolog clusters.

Figure S4. Numbers of gene duplications mapped across a non-binary species tree as estimated by Notung. The topology used is the rosid portion of the ML topology of the nuclear

concatenated alignment of 1,103 genes, with poorly supported relationships collapsed. duplications were counted from 8,324 homolog clusters.

Figure S5. Root-to-tip lengths per taxon with partitions of fixed local clocks indicated. Pruned taxa with outlier root-to-tip lengths are indicated with an X, partitions are indicated with colors. (a) FLC3, (b) FLC6, (c) FLC8.

Figure S6. Chronogram estimated under the UCLN clock model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S7. Chronogram estimated under the UCLN clock model, with alternative prior 2. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S8. Chronogram estimated under the RLC model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S9. Chronogram estimated under the FLC3 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S10. Chronogram estimated under the FLC6 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S11. Chronogram estimated under the FLC8 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S12. Chronogram estimated under the FLC8 model, with alternative prior 1. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles, with alternative calibrations as red circles.

Figure S13. Chronogram estimated under the STRC model. Numbers behind nodes indicate 95% HPD intervals. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S14. Substitution rates as estimated in FLC8 analyses for the different clock partitions. Boxplots for each partition for (a) alternative prior 1 and (b) the “normal” prior setting. Colors correspond to the partitions as shown in Figures 5, S14, S15 and S18.

Figure S15. Histograms of age estimates of duplication nodes, for (a) the duplications mapped to the legume crown node in the Notung analysis and for duplication nodes in gene trees with only (b) Detarioideae, (c) Caesalpinioideae and (d) Papilionoideae included.

Table S1. Taxon occupancy per analysis (number of gene trees | number of sequences)

Species	Phyparts (n=8,038)	Notung (n=8,324)	WGDgc (n=8,324)	GRAMPA (n=4,731)	species tree dating supermatrix (n=36)	legume WGD dating (n=863)	Caesalpinioideae WGD dating (n=246)	Detarioideae WGD dating (n=250)	Papilionoideae WGD dating (n=272)
Acacia koa	6917 | 8314	7256 | 8061			35	783 | 1000			
Acrocarpus fraxinifolius	5002 | 5765	5141 | 5536			27	582 | 678			
Afzelia bella	5165 | 6074	5371 | 5929	5371 | 5929		30	565 | 699		220 | 288	
Albizia julibrissin	6905 | 8980	7183 | 8544	7183 | 8544	3914 | 4548	30	793 | 1142	237 | 400		
Alnus serrulata	5253 | 5768	3633 | 3667	3633 | 3667	1389 | 1390	28	572 | 577	126 | 126	145 | 145	145 | 145
Amaranthus hypochondriacus	3582 | 3725				29				
Anthonotha fragrans	6413 | 8046	6740 | 7793	6740 | 7793	3727 | 3862	35	704 | 932		250 | 391	
Apios americana	7278 | 9576	7711 | 9249	7711 | 9249		35	809 | 1138			
Aquilegia coerulea					34				
Arabidopsis thaliana	4278 | 4537	1706 | 1724				249 | 249			
Arachis ipaensis	6679 | 8933	6748 | 8204	6748 | 8204		35	724 | 1002			272 | 365
Astragalus membranaceus	6104 | 7190	6521 | 7152			31				
Bauhinia tomentosa	6428 | 7153	6745 | 6907	6745 | 6907	3696 | 3727	35	712 | 731			
Bituminaria bituminosa	5810 | 6986	6142 | 6861			30				
Cajanus cajan	7393 | 11163	7654 | 10264	7654 | 10264		33				
Cannabis sativa	6606 | 7331	4424 | 4462	4424 | 4462	1580 | 1589	36	663 | 665	138 | 139	133 | 139	157 | 157
Carica papaya	5368 | 5900	2882 | 2897			26	476 | 478			
Castanea mollissima	6600 | 7797	4683 | 4783	4683 | 4783	1765 | 1777	33	685 | 695	149 | 151	153 | 155	171 | 172
Cercis canadensis	4449 | 4925	4579 | 4655	4579 | 4655		24	498 | 510			
Chamaecrista fasciculata	2206 | 2374	2100 | 2165			11				
Cicer arietinum	7237 | 9901	7508 | 9294	7508 | 9294		36	806 | 1190			
Citrus sinensis	6474 | 7320	3675 | 3704	3675 | 3704		34	598 | 600			
Cladrastis lutea	3779 | 4244	3867 | 4115			19				
Codariocalyx motorius	6008 | 7305	6379 | 7240			33				
Copaifera officinalis	5750 | 6983	6075 | 6849	6075 | 6849		29	658 | 842		240 | 365	
Cucumis sativus	5749 | 6290	3236 | 3280	3236 | 3280	932 | 933	36	484 | 486	86 | 88	79 | 81	118 | 121
Desmanthus illinoensis	1714 | 1919	1614 | 1718			8				
Elaeocarpus photiniifolia	1067 | 1144	435 | 440			4	92 | 92			
Entada abyssinica	6799 | 8535	7112 | 8193	7112 | 8193	3919 | 4435	33	790 | 1109	246 | 384		
Eucalyptus grandis	6224 | 7061	2946 | 3012			36	468 | 476			
Fragaria vesca	5752 | 6532	3949 | 4032	3949 | 4032	1473 | 1485	28	605 | 610	131 | 132	130 | 131	135 | 136
Gleditsia triacanthos	4090 | 4809	4234 | 4627			21	418 | 502			

Glycine max	7871 | 12902	8211 | 11828	8211 | 11828		36	863 | 1551			272 | 525
Glycyrrhiza lepidota	6287 | 7531	6657 | 7399			32				
Gossypium raimondii	7198 | 8632	4309 | 4445	4309 | 4445		36	684 | 703			
Gymnocladus dioicus	1764 | 2002	1827 | 1956			4				
Inga spectabilis	7029 | 8767	7405 | 8534	7405 | 8534	4092 | 4650	35	796 | 1115	239 | 394		
Juglans regia	6882 | 8080	4953 | 5078	4953 | 5078	1915 | 1928	36	736 | 749	171 | 173	180 | 185	188 | 190
Lactuca sativa	4808 | 5134				33				
Lathyrus sativus	5628 | 6723	5923 | 6628			27				
Lens culinaris	2061 | 2339	2001 | 2170			12				
Linum usitatissimum	5136 | 5472	2330 | 2337				360 | 361			
Lotus japonicus	6656 | 8793	6864 | 8210	6864 | 8210		34	766 | 1039			
Lupinus angustifolius	4102 | 4612	4232 | 4503			19				
Lupinus polyphyllus	1962 | 2194	1919 | 2041			6				
Manihot esculenta	7119 | 8412	4219 | 4312	4219 | 4312		35	675 | 682			
Medicago truncatula	7634 | 11190	7952 | 10426	7952 | 10426	4240 | 4545	36	841 | 1311			269 | 464
Microlobius foetidus	7074 | 9088	7305 | 8596	7305 | 8596	3966 | 4583	33	799 | 1140	235 | 396		
Mimulus guttatus	5367 | 5767				34				
Morus notabilis	6386 | 7223	4292 | 4340	4292 | 4340	1507 | 1516	36	665 | 667	140 | 141	131 | 137	150 | 150
Nelumbo nucifera	4957 | 5050				35				
Paeonia lactiflora	2888 | 3096				14				
Panax ginseng	2198 | 2333								
Papaver somniferum					32				
Phaseolus vulgaris	7746 | 11727	8089 | 10882	8089 | 10882		36	857 | 1411			271 | 485
Pisum sativum	6251 | 7822	6554 | 7556			30				
Polygala lutea	4072 | 4349	3074 | 3108	3074 | 3108	1335 | 1339		366 | 366	68 | 69	76 | 78	90 | 92
Populus trichocarpa	7134 | 11983	4190 | 6196	4190 | 6196		36	664 | 1062			
Primula veris	3765 | 3997				29				
Prioria balsamifera	6287 | 7625	6646 | 7485	6646 | 7485		35	686 | 886		244 | 361	
Prosopis alba	2799 | 3274	2790 | 3069			11				
Prunus persica	7260 | 8487	5045 | 5156	5045 | 5156	1823 | 1837	36	764 | 772	162 | 164	150 | 153	168 | 169
Punica granatum	3825 | 4205	1649 | 1684			21	283 | 289			
Quillaja saponaria	4680 | 5024	4214 | 4239	4214 | 4239	2114 | 2117	24	515 | 516	145 | 145	164 | 164	145 | 145
Salix purpurea	7058 | 11780	4128 | 6094	4128 | 6094		35	655 | 1050			
Senna hebecarpa	1643 | 1835	1573 | 1658			6				
Solanum tuberosum	5070 | 5529				28				
Styphnolobium japonicum	6964 | 7781	7233 | 7421			34	757 | 780			

Theobroma cacao	7156 | 8588	4304 | 4428	4304 | 4428		36	683 | 701
Trifolium pratense	4936 | 5893	5126 | 5729			19	
Tripterygium wilfordii	4551 | 4847	2446 | 2458			25	380 | 382
Vicia faba	4277 | 5054	4350 | 4833			25	
Vigna radiata	6526 | 9119	6726 | 8500	6726 | 8500		33	
Vitis vinifera	5768 | 6576				32	
Xanthocercis zambesiaca	6496 | 7518	6934 | 7444			34	716 | 813
Zenia insignis	5442 | 6012	5569 | 5684	5569 | 5684	2824 | 2868	25	610 | 616

Table S2. Age intervals specified for the fossil calibration priors under different alternative priors.

Calibration	Definition	MRCA taxon 1	MRCA taxon 2	Prior	Alternative prior 1	Alternative prior 2
eudicots						
26	CG eudicots	*Aquilegia coerulea*	*Medicago truncatula*	normal (mean 126.0, stdev 1.0)	normal (mean 126.0, stdev 1.0)	normal (mean 126.0, stdev 1.0)
27	CG Ranunculales	*Aquilegia coerulea*	*Papaver somniferum*	uniform (min 113.0, max 126.0)	uniform (min 113.0, max 126.0)	uniform (min 113.0, max 126.0)
38	CG Pentapetalae	*Nelumbo nucifera*	*Medicago truncatula*	uniform (min 100.0, max 126.0)	uniform (min 100.0, max 126.0)	uniform (min 100.0, max 126.0)
48	SG Ericales	*Primula veris*	*Solanum tuberosum*	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 113.0)
94	SG Myrtaceae	*Eucalyptus grandis*	*Punica granatum*	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 100.0)
105	SG Brassicales	*Carica papaya*	*Theobroma cacao*	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 100.0)
112	CG Rosaceae	*Fragaria vesca*	*Prunus persica*	uniform (min 49.4, max 126.0)	uniform (min 49.4, max 126.0)	uniform (min 49.4, max 66.0)
116	SG Cannabaceae	*Cannabis sativa*	*Morus notabilis*	uniform (min 66.0, max 126.0)	uniform (min 66.0, max 126.0)	uniform (min 66.0, max 83.6)
122	SG Juglandaceae	*Alnus serrulata*	*Juglans regia*	uniform (min 64.4, max 126.0)	uniform (min 64.4, max 126.0)	uniform (min 64.4, max 83.6)
133	SG Populus	*Populus trichocarpa*	*Salix purpurea*	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 56.0)
X14	SG Fagales	*Alnus serrulata*	*Medicago truncatula*	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 126.0)
legumes						
A	SG Leguminosae	*Medicago truncatula*	*Quillaja saponaria*	uniform	uniform	uniform

				(min 63.5, max 126.0)	(min 63.5, max 126.0)	(min 63.5, max 100.0)
C	SG Cercis	*Cercis canadensis*	*Bauhinia tomentosa*	uniform (min 36.0, max 126.0)		uniform (min 36.0, max 83.6)
C&	SG Bauhinia	*Bauhinia tomentosa*	*Cercis canadensis*		uniform (min 46.0, max 126.0)	
F	CG Resin-producing clade	*Copaifera officinalis*	*Prioria balsamifera*	uniform (min 22.5, max 126.0)	uniform (min 22.5, max 126.0)	uniform (min 22.5, max 66.0)
G	SG Detarioideae	*Copaifera officinalis*	*Bauhinia tomentosa*	uniform (min 53.0, max 126.0)		uniform (min 53.0, max 83.6)
G&	SG Resin-producing clade	*Copaifera officinalis*	*Anthonotha fragrans*		uniform (min 53.0, max 126.0)	
H&	CG Amherstieae	*Afzelia bella*	*Anthonotha fragrans*		uniform (min 46.0, max 126.0)	
I2	SG Styphnolobium/Cladrastis	*Styphnolobium japonicum*	*Medicago truncatula*	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 66.0)
M2	SG Robinioid clade	*Lotus japonicus*	*Medicago truncatula*	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 66.0)
Q	SG Acacieae/Ingeae	*Albizia julibrissin*	*Prosopis alba*	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 66.0)
Q2	SG Acacia s.s.	*Acacia koa*	*Albizia julibrissin*	uniform (min 23.0, max 126.0)	uniform (min 23.0, max 126.0)	uniform (min 23.0, max 56.0)
Z	SG Caesalpinioideae	*Albizia julibrissin*	*Medicago truncatula*	uniform (min 58.0, max 126.0)	uniform (min 58.0, max 126.0)	uniform (min 58.0, max 83.6)

Table S3. Node age estimates and priors (95% HPD intervals) of nodes A-H in the different analyses.

Node	A	B	C	D	E	F	G	H
Clade	**Leguminosae**	**Cerc+Detar**	**Cercidoideae**	**Detarioideae**	**Dial+Caes+Pap**	**Caes+Pap**	**Caesalpinioideae**	**Papilionoideae**
Standard prior								
Marginal prior	79.37 - 109.20	54.56 - 99.48	36.00 - 80.55	28.91 - 87.21	73.77 - 106.04	68.16 - 101.69	56.31 - 95.76	58.85 - 96.39
UCLN	65.47 - 86.45	57.50 - 80.75	36.00 - 53.97	25.47 - 42.98	63.51 - 84.73	60.64 - 81.67	54.11 - 74.49	55.19 - 73.58
RLC	73.46 - 81.18	68.06 - 75.69	39.34 - 46.74	31.52 - 36.43	69.77 - 77.35	68.05 - 75.45	55.76 - 63.75	49.05 - 54.38
strict clock	66.94 - 69.55	60.45 - 63.87	36.00 - 36.66	26.25 - 28.71	65.60 - 68.22	64.90 - 67.48	56.01 - 59.12	56.89 - 59.47
FLC 3 clocks	65.99 - 68.85	60.79 - 64.20	36.00 - 36.85	27.69 - 30.59	63.77 - 66.52	62.78 - 65.43	56.09 - 59.04	47.39 - 50.03
FLC 6 clocks	65.74 - 68.81	60.70 - 64.40	36.00 - 36.86	27.53 - 30.94	63.53 - 66.47	62.57 - 65.43	56.10 - 59.20	47.24 - 49.86
FLC 8 clocks	64.63 - 67.64	60.24 - 64.79	36.00 - 52.41	27.00 - 49.18	62.72 - 65.61	61.86 - 64.65	55.53 - 58.51	46.98 - 49.60
Alternative prior 1 (Bruneau et al. 2008 for Cercidoideae and Detarioideae)								
Marginal prior	81.60 - 110.20	63.87 - 103.63	46.00 - 85.81	53.00 - 90.89	75.00 - 106.55	69.90 - 103.35	57.41 - 97.38	60.69 - 98.39
FLC 8 clocks	64.81 - 67.96	64.06 - 67.35	57.35 - 63.89	55.38 - 63.49	63.08 - 65.94	62.16 - 64.90	55.73 - 58.76	46.96 - 49.59
Alternative prior 2 (tighter maxima)								
Marginal prior	73.30 - 96.56	56.01 - 83.60	36.00 - 71.11	28.75 - 75.30	66.75 - 90.79	64.04 - 83.60	53.37 - 81.05	54.12 - 79.22
UCLN	66.92 - 76.45	60.43 - 72.34	36.01 - 50.55	39.85 - 52.84	63.48 - 71.28	61.45 - 69.38	54.20 - 63.50	47.59 - 55.25
Alternative prior 3 (reduced taxon sampling)								
Marginal prior	72.85 - 106.32	53.45 - 95.52	36.00 - 78.70	29.28 - 85.30	64.01 - 100.04	58.03 - 91.93	38.95 - 83.49	46.50 - 86.28
UCLN	64.72 - 79.69	57.35 - 75.43	36.00 - 53.18	25.40 - 39.20	62.29 - 76.62	60.08 - 74.12	49.28 - 66.70	49.68 - 64.63

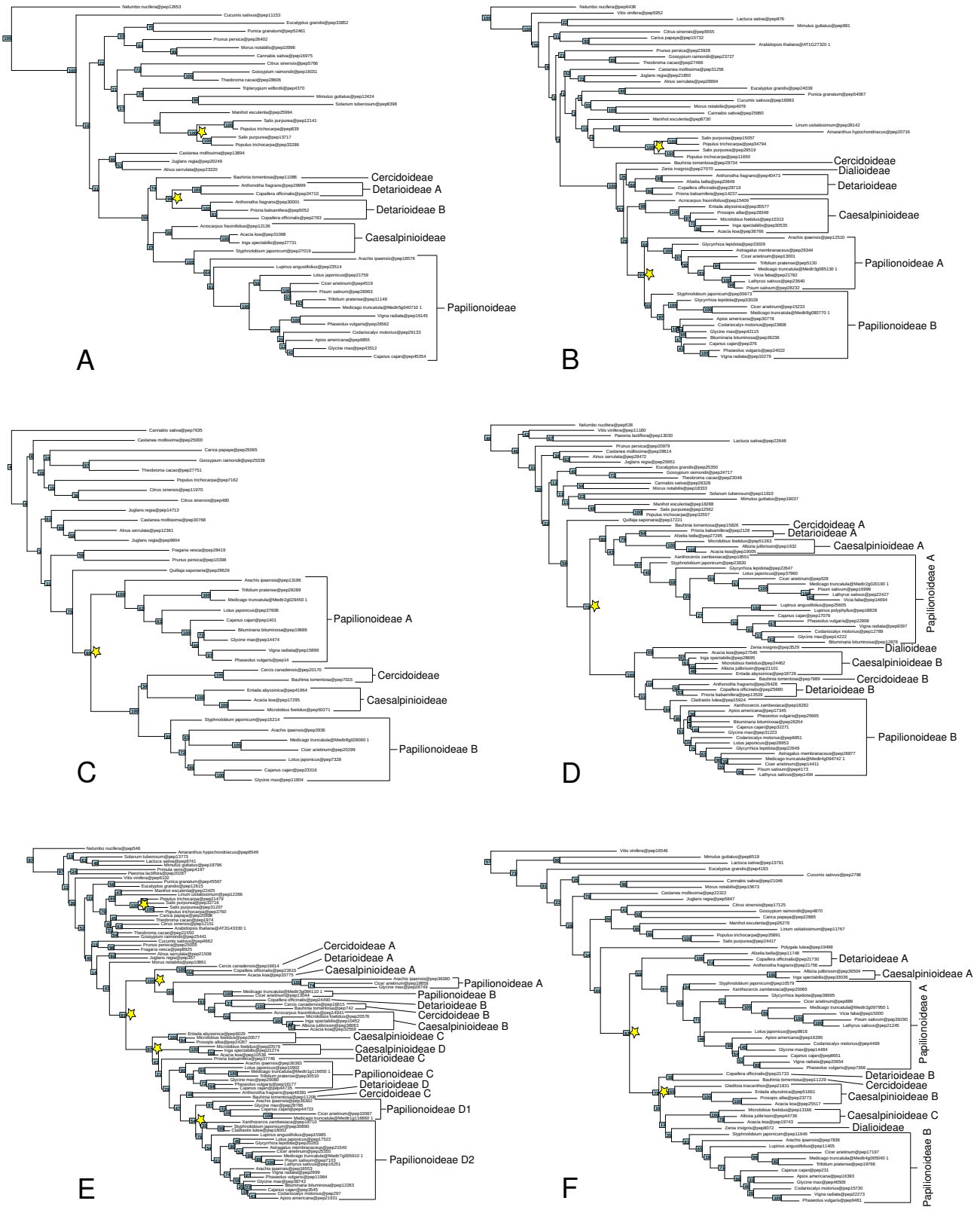


Figure S1. Examples of homolog clusters with gene duplications in legumes that passed the bootstrap filter. Yellow stars behind nodes indicate locations of gene duplications, numbers on nodes indicate bootstrap support. The plotted gene trees are extracted from (A) cluster3675_1rr_1rr, showing a duplication subtending Detarioideae, (B) cluster1032_1rr_1rr, showing a duplication subtending Papilionoideae, (C) cluster1248_1rr_1rr and (D) cluster2941_1rr_1rr, both with a duplication subtending the legume family. Trees for (E) cluster51_7rr_1rr and (F) cluster544_1rr_1rr show evidence of more than one duplication, including one specific to Papilionoideae in the former.

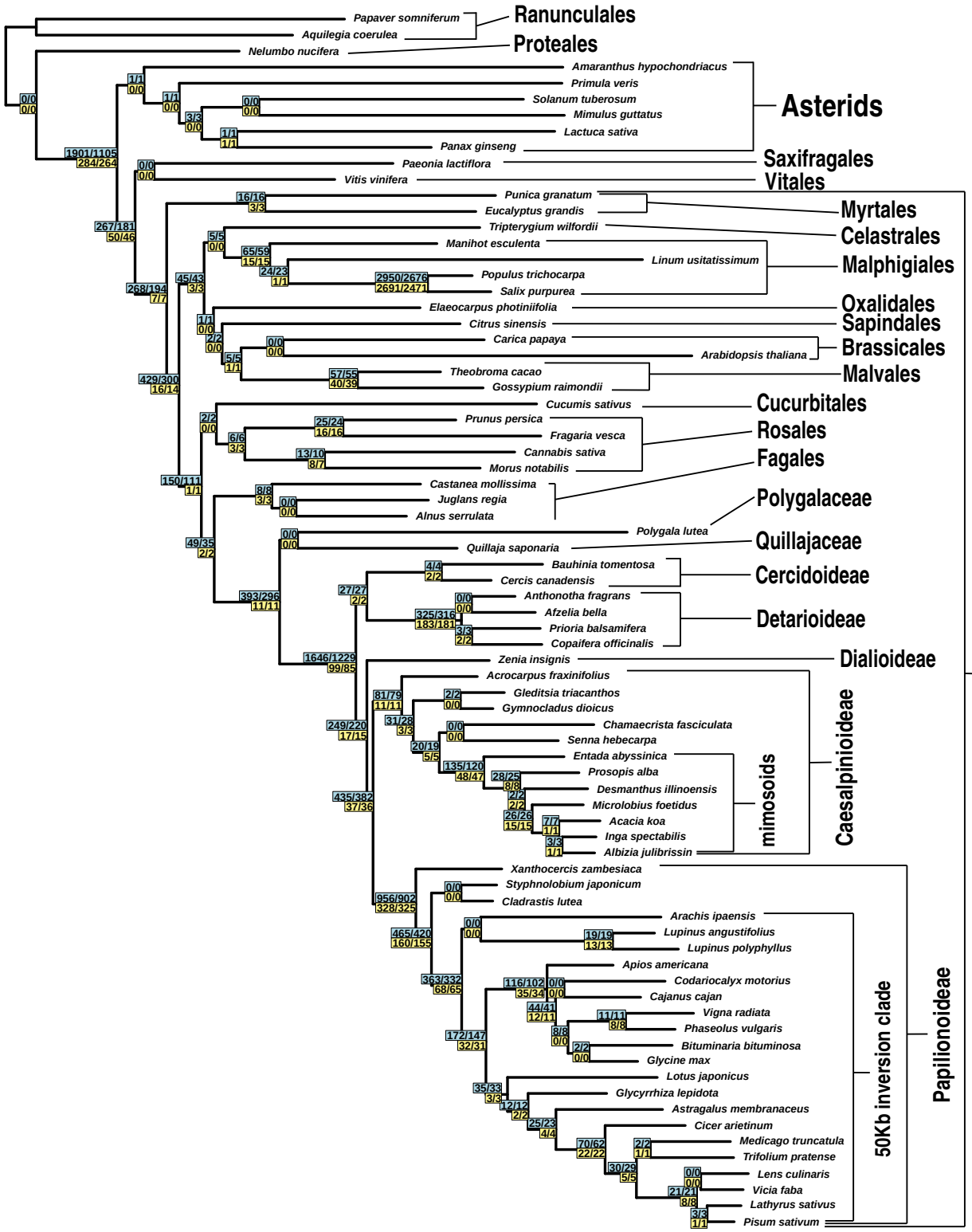


Figure S2. Numbers of gene duplications mapped across the phylogeny with Phyparts. The topology used is the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,038 homolog clusters. Numbers above branches (with blue background) and below branches (with yellow background) represent numbers of duplications and numbers of homolog trees with duplications, without or with a bootstrap filter of 50%, respectively.

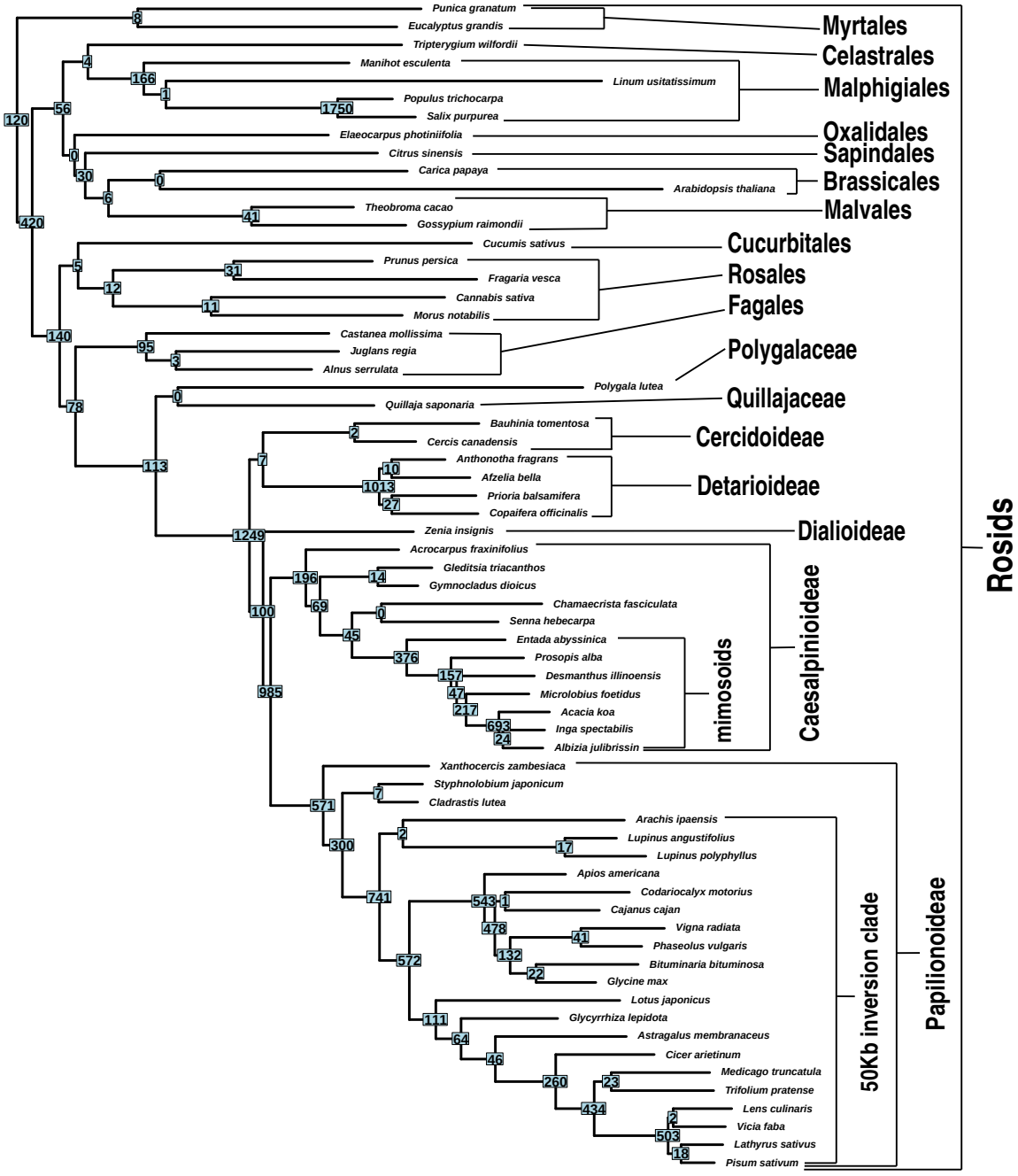


Figure S3. Numbers of gene duplications mapped across the species tree as estimated by Notung. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,324 homolog clusters.

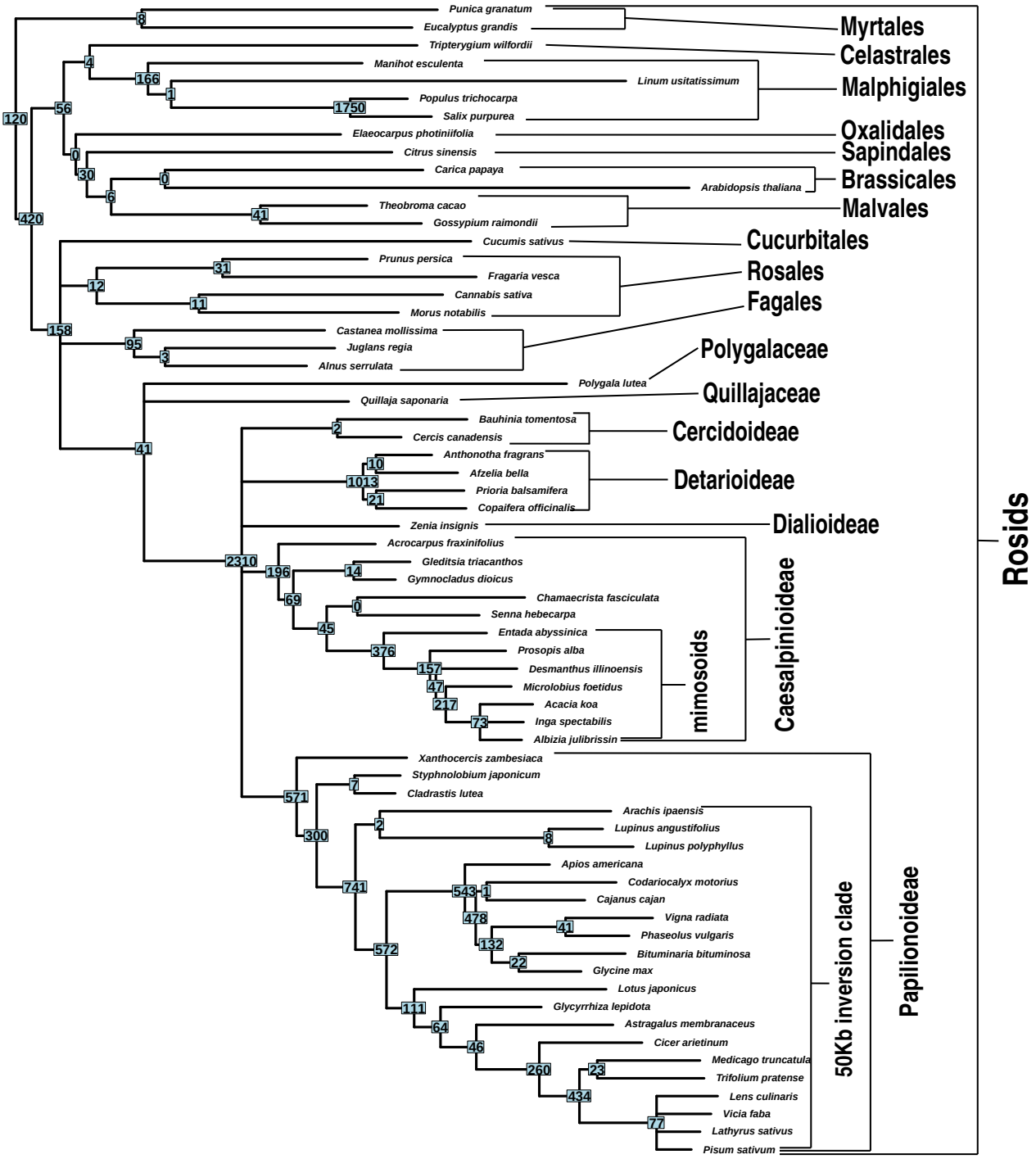


Figure S4. Numbers of gene duplications mapped across a non-binary species tree as estimated by Notung. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes, with poorly supported relationships collapsed. duplications were counted from 8,324 homolog clusters.

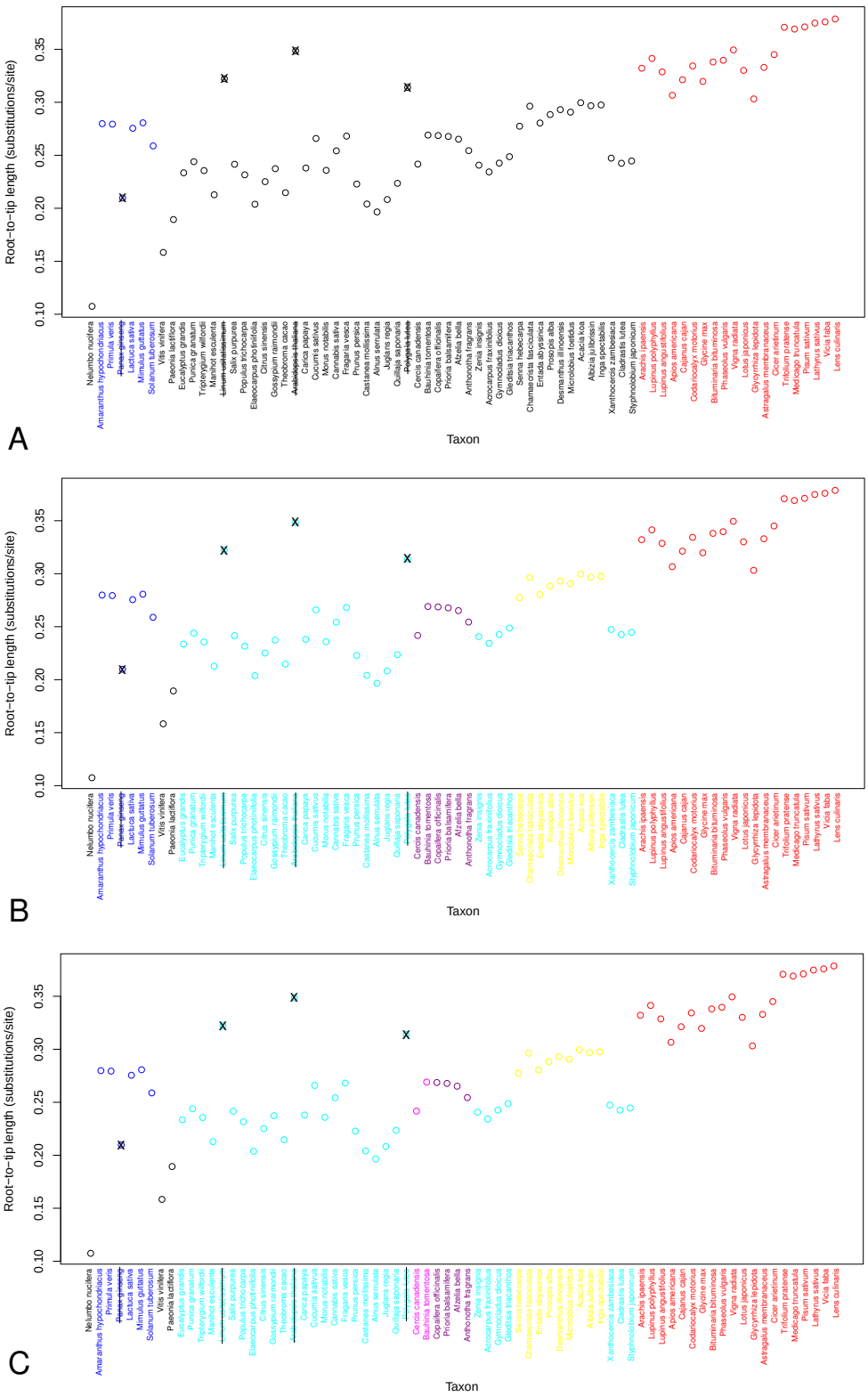


Figure S5. Root-to-tip lengths per taxon with partitions of fixed local clocks indicated. Pruned taxa with outlier root-to-tip lengths are indicated with an X, partitions are indicated with colors. (A) FLC3, (B) FLC6, (C) FLC8.

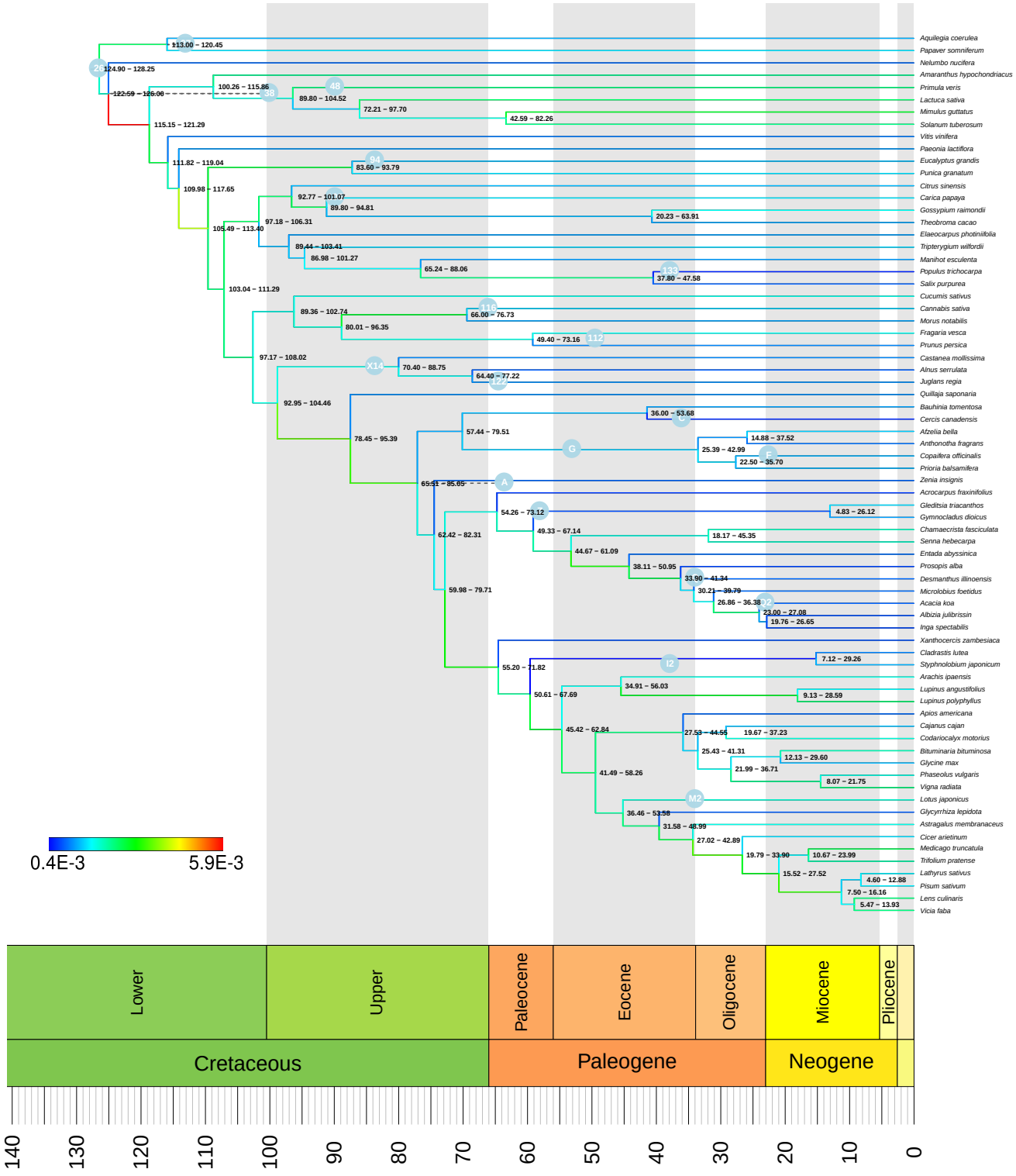


Figure S6. Chronogram estimated under the UCLN clock model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

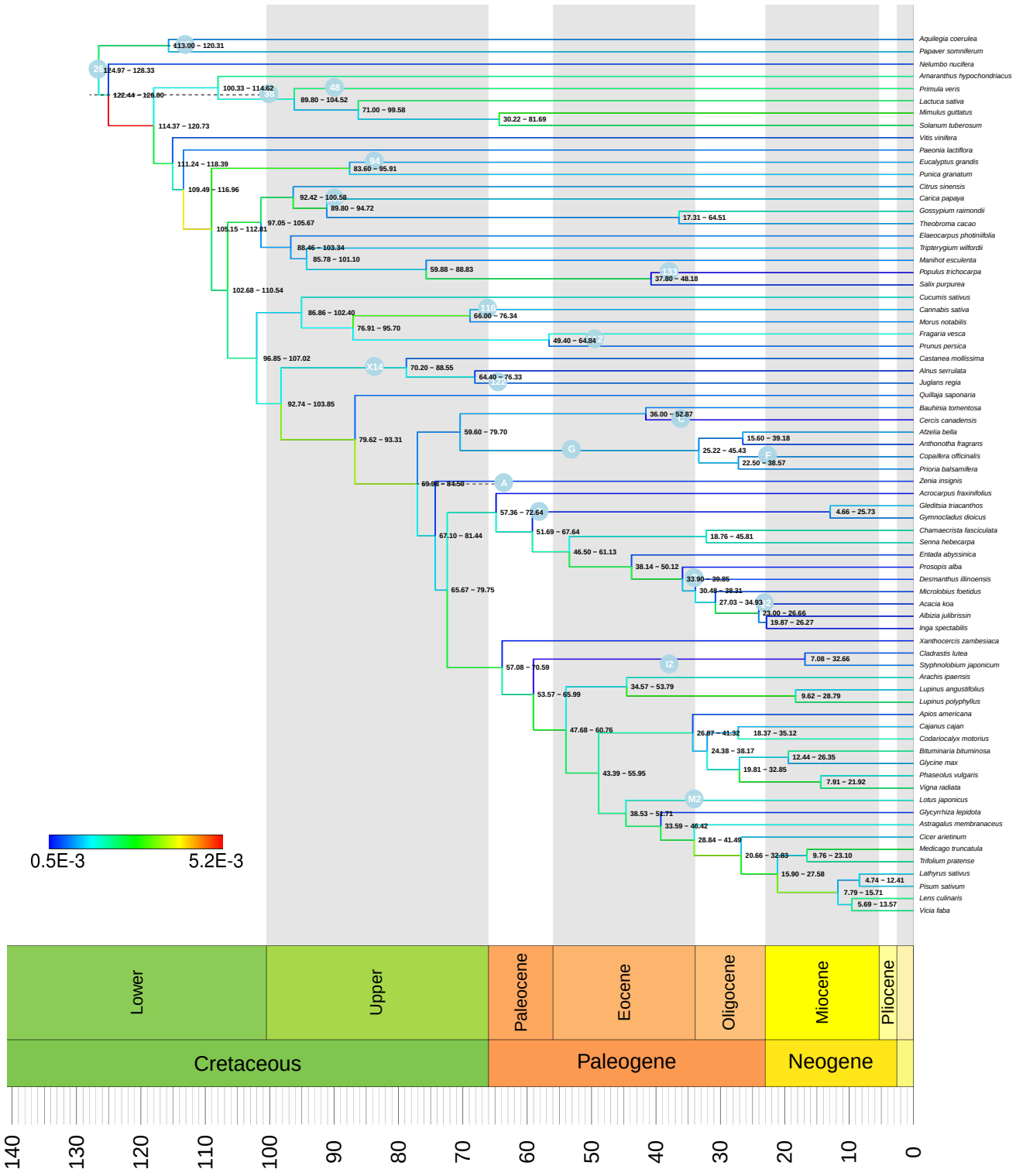


Figure S7. Chronogram estimated under the UCLN clock model, with alternative prior 2. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

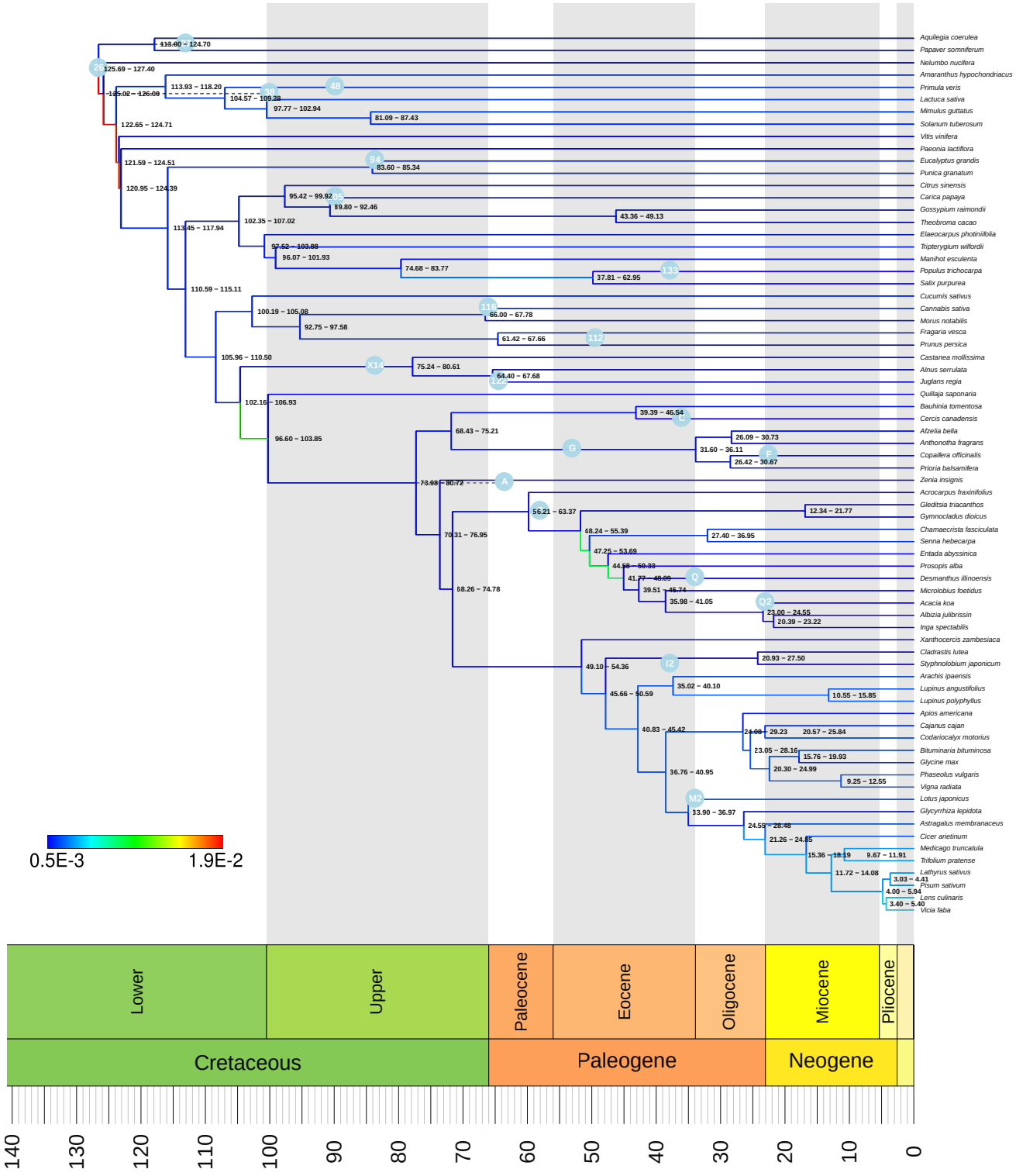


Figure S8. Chronogram estimated under the RLC model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as indicated by the color legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

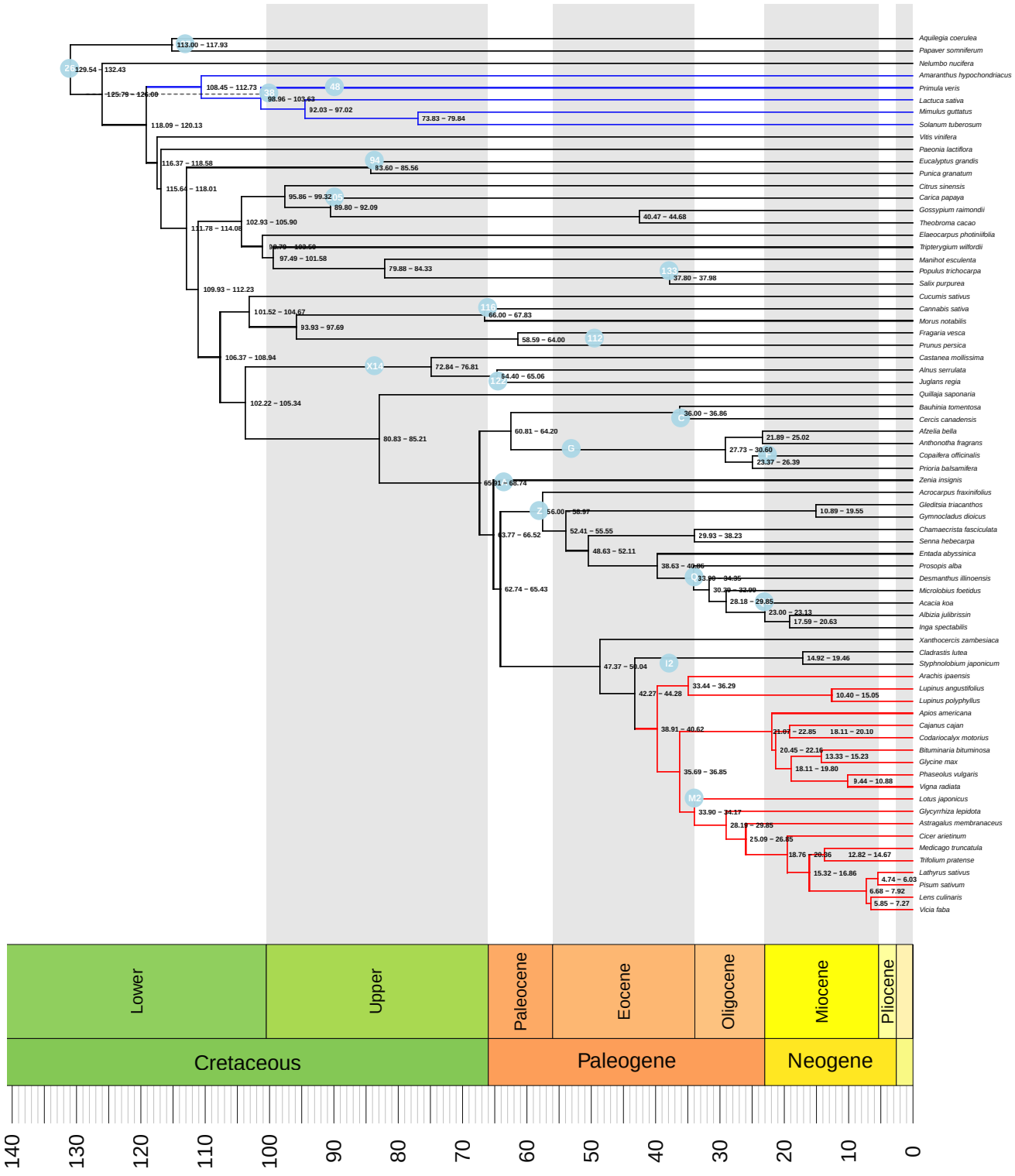


Figure S9. Chronogram estimated under the FLC3 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

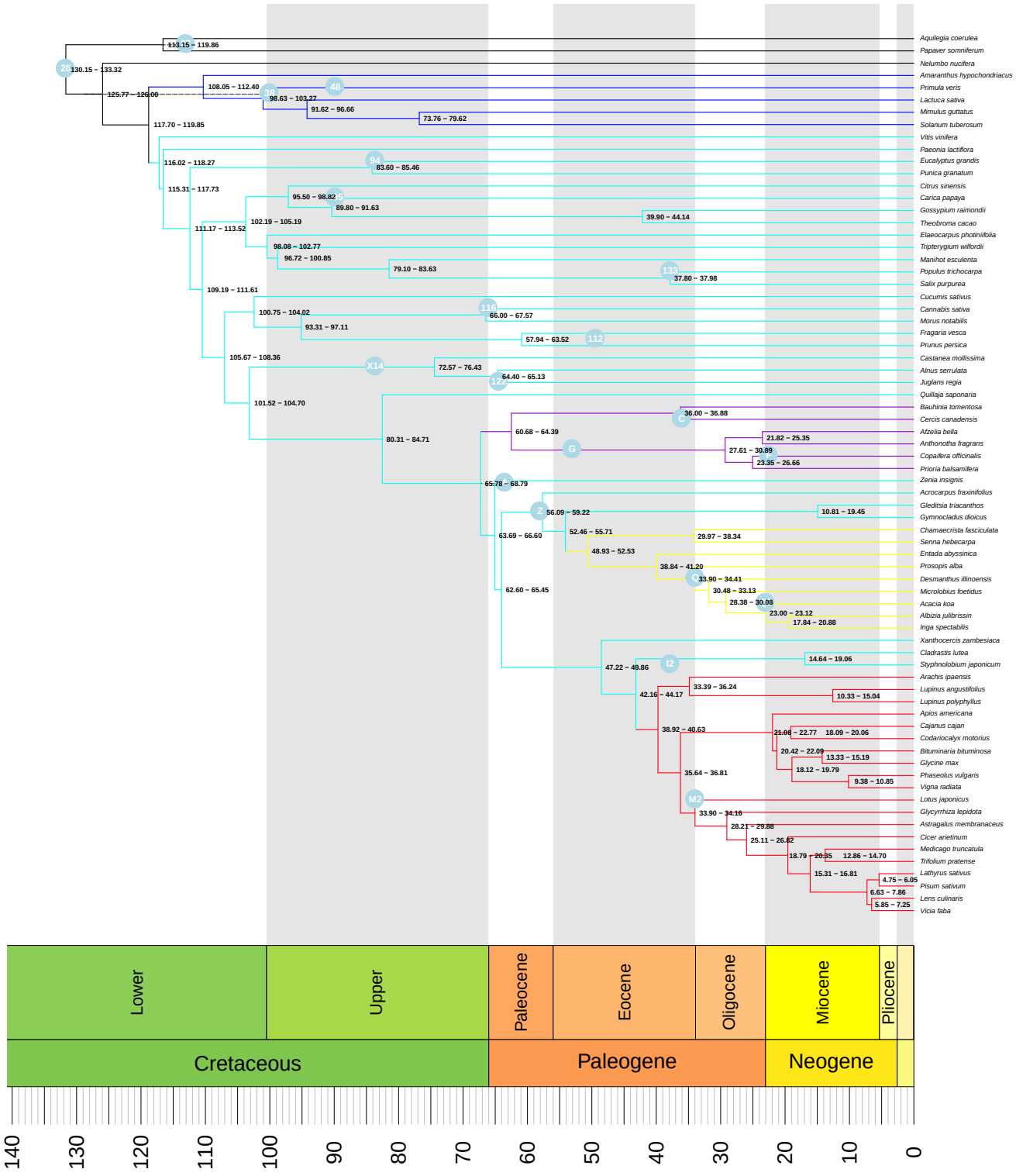


Figure S10. Chronogram estimated under the FLC6 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

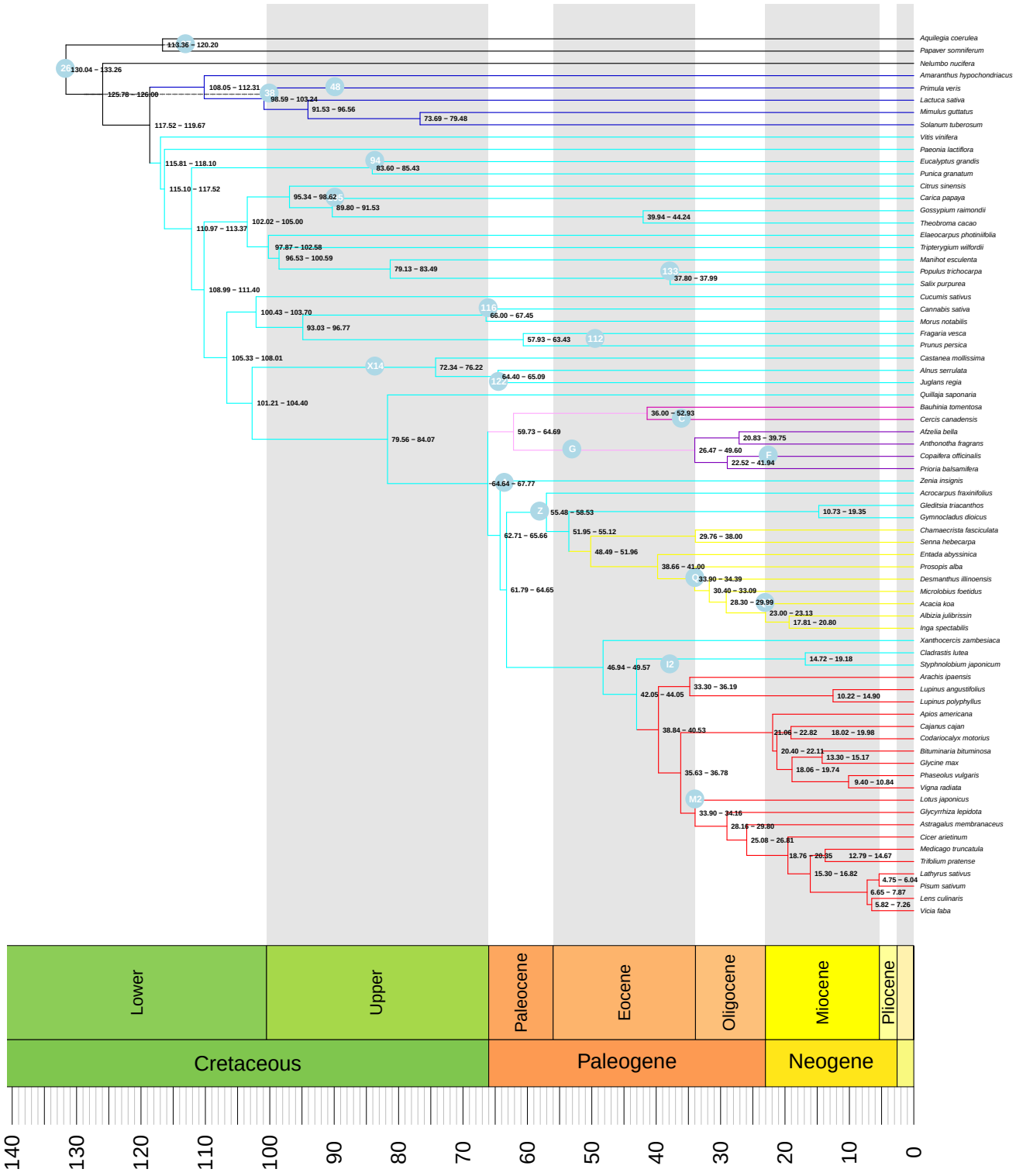


Figure S11. Chronogram estimated under the FLC8 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

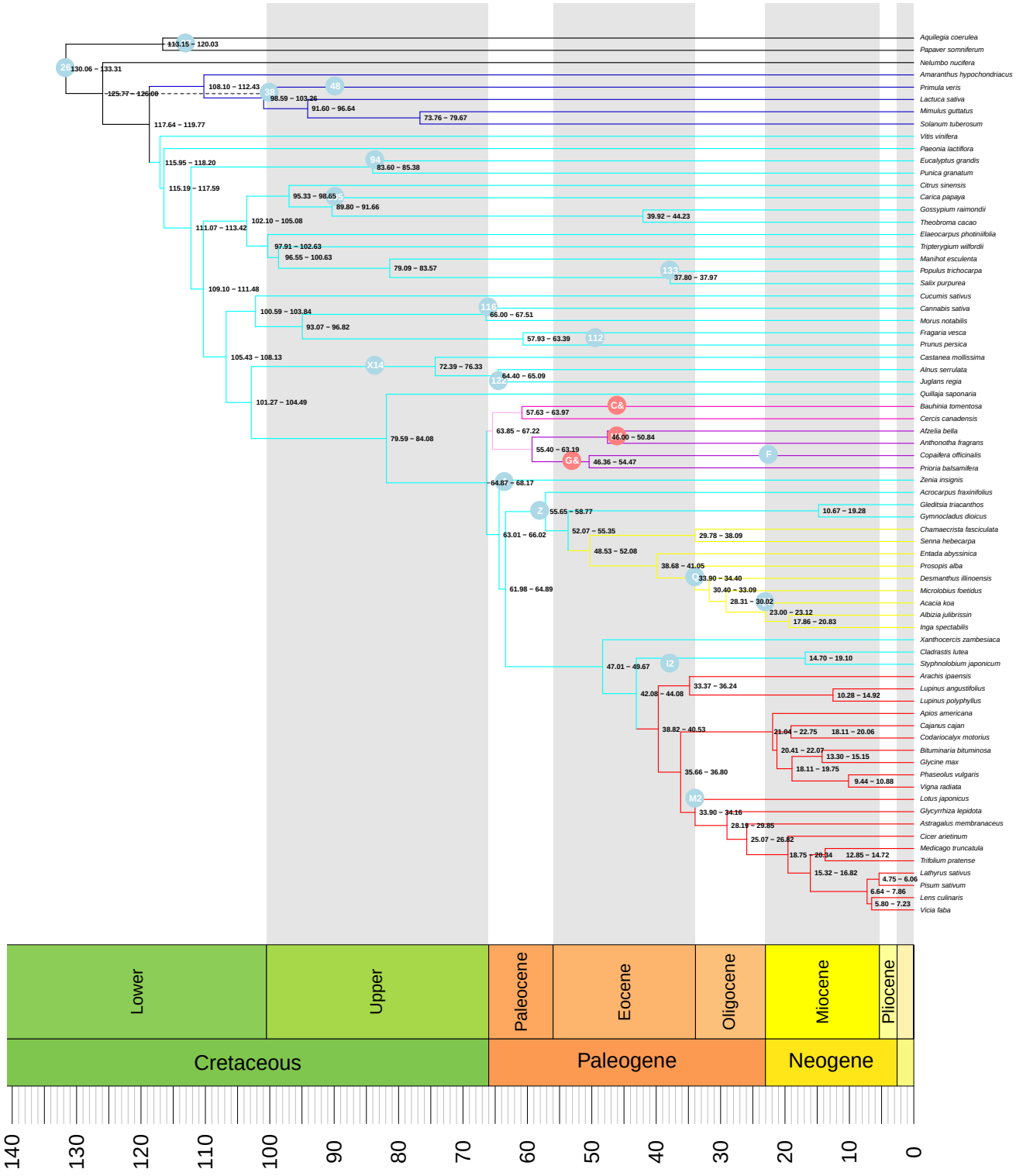


Figure S12. Chronogram estimated under the FLC8 model, with alternative prior 1. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles, with alternative calibrations as red circles.

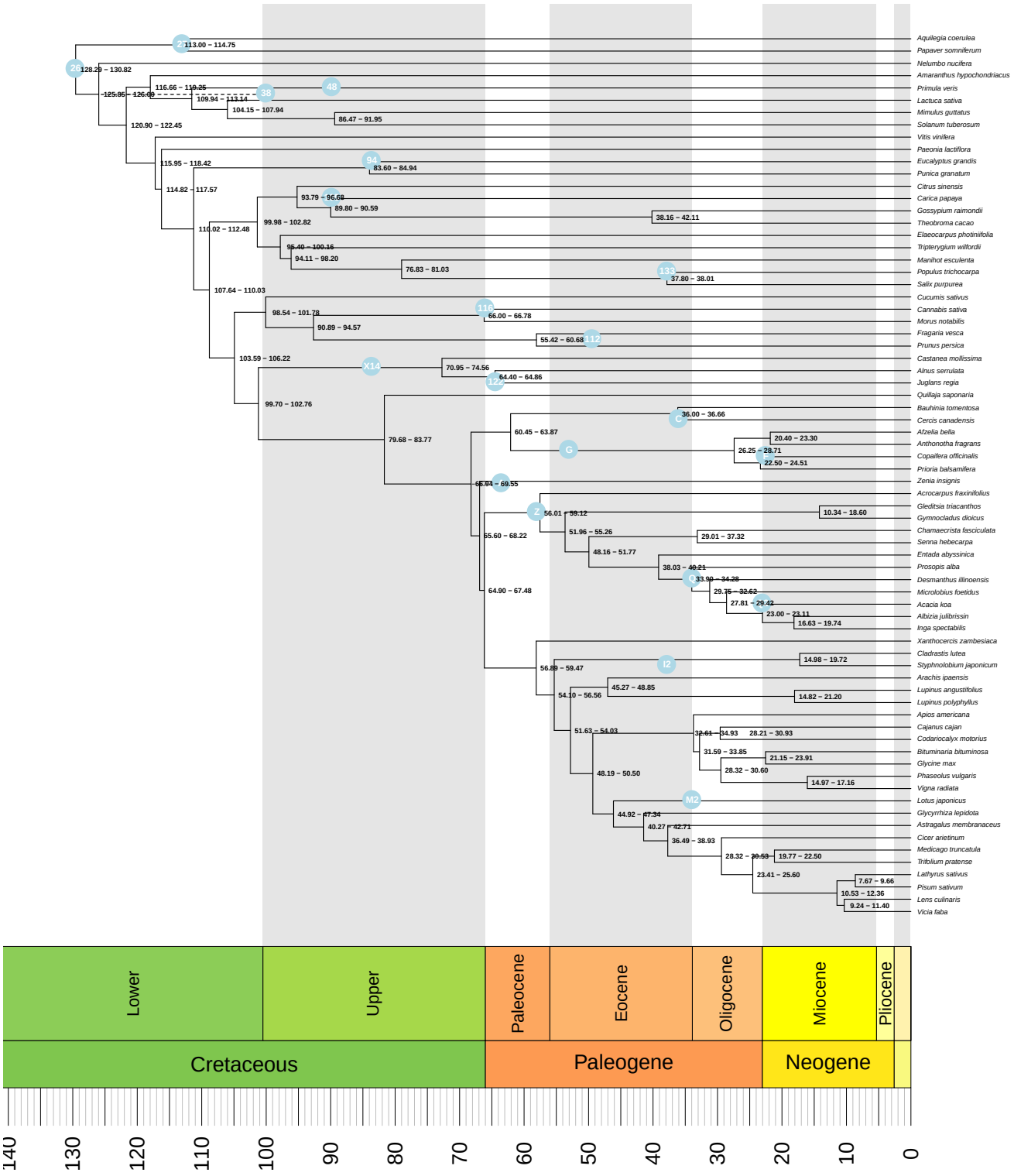
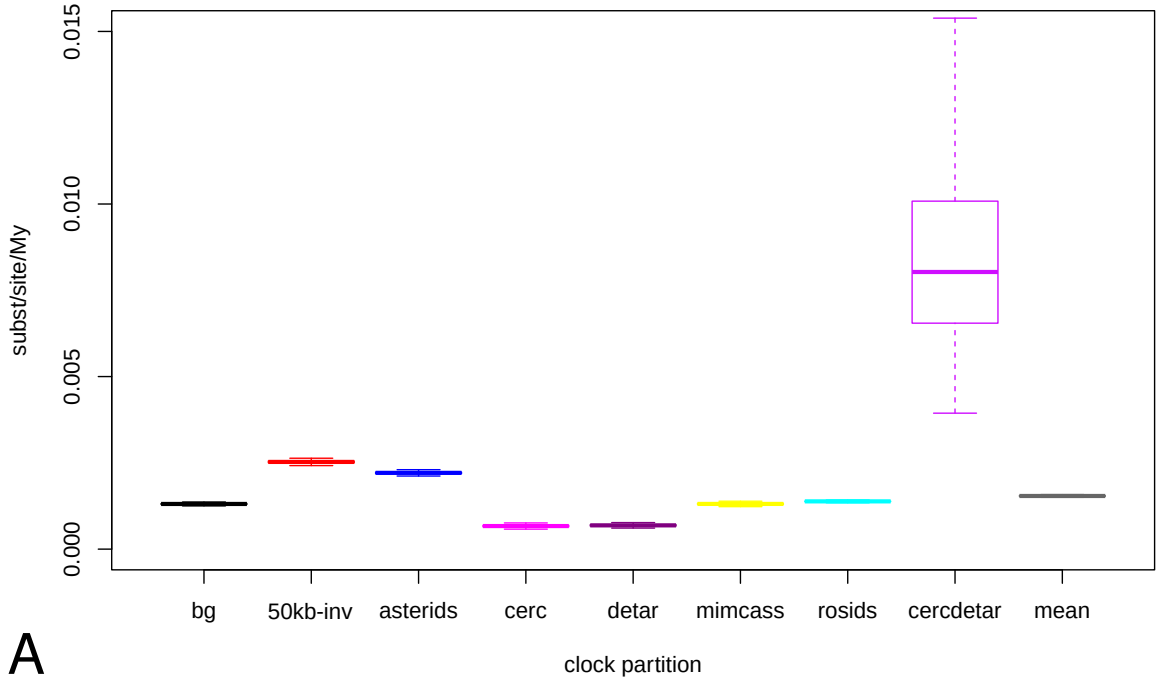


Figure S13. Chronogram estimated under the STRC model. Numbers behind nodes indicate 95% HPD intervals. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.



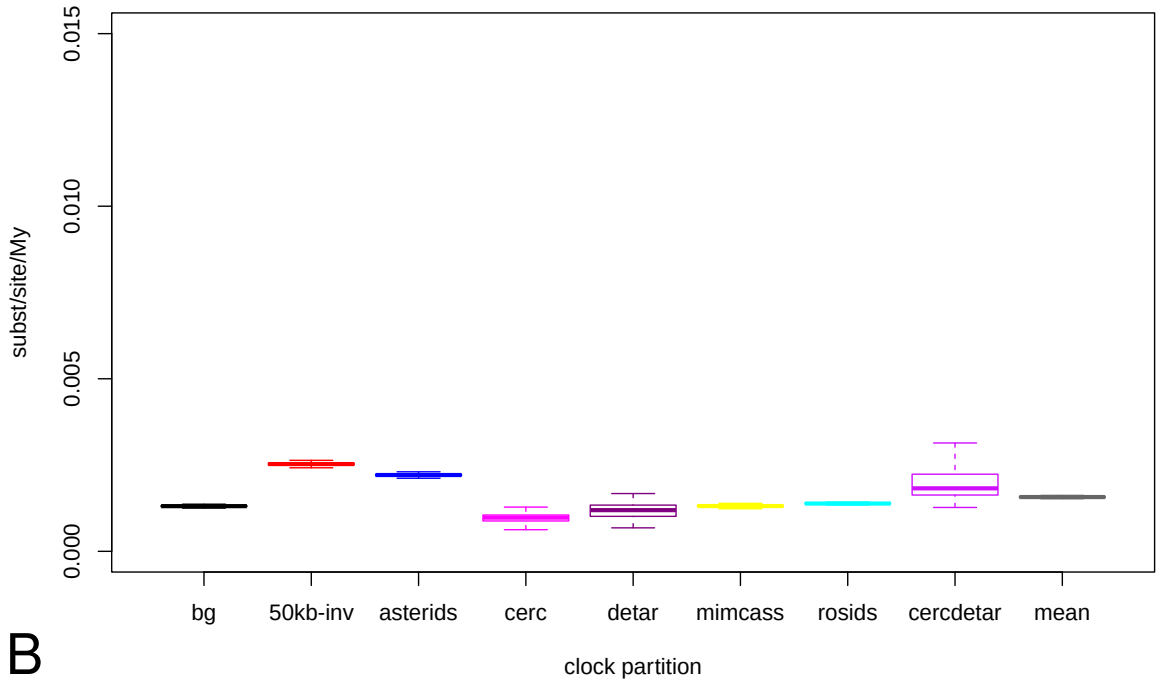


Figure S14. Substitution rates as estimated in FLC8 analyses for the different clock partitions. Boxplots for each partition for (A) alternative prior 1 and (B) the “normal” prior setting. Colors correspond to the partitions as shown in Figs 5, S14, S15 and S18.

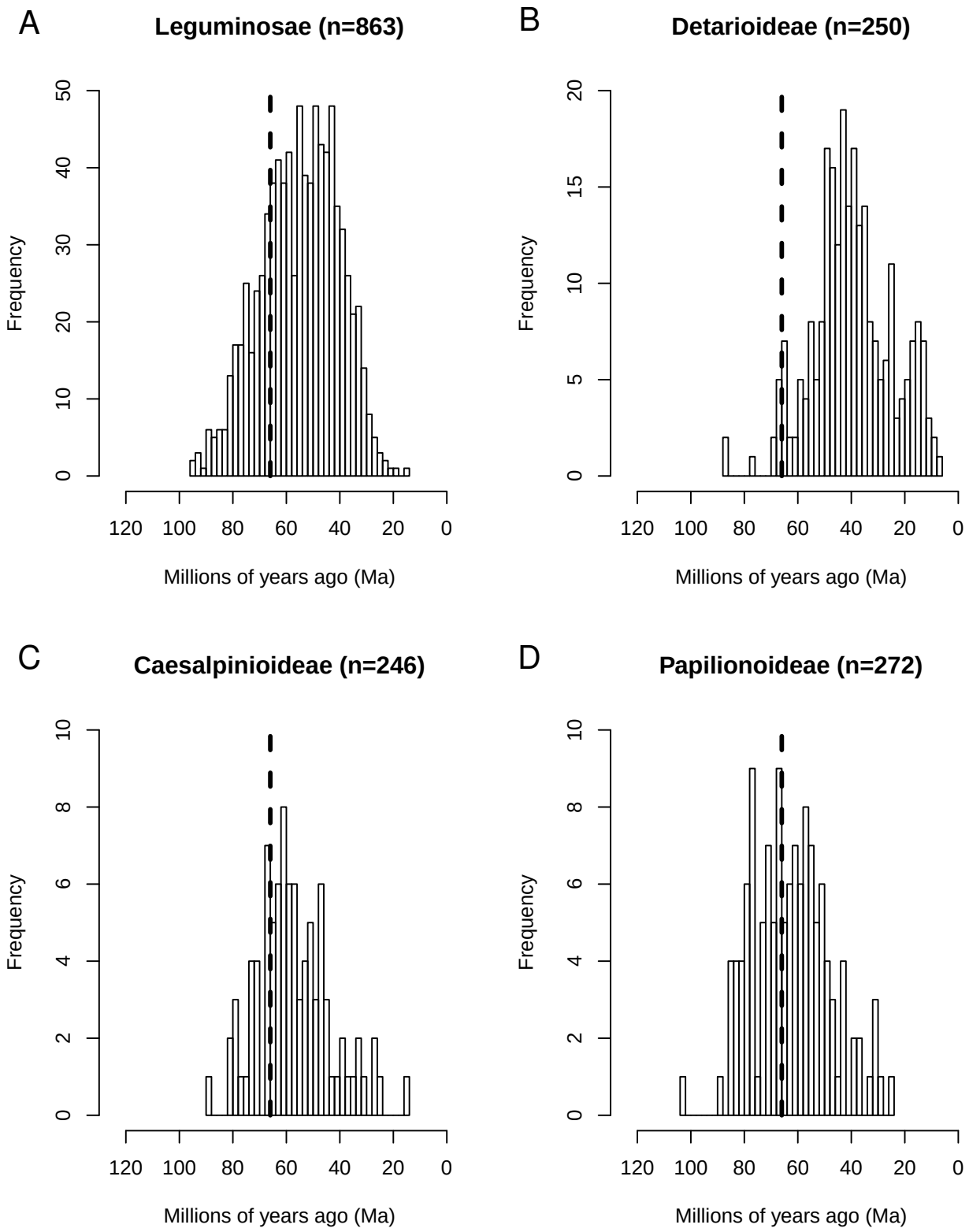


Figure S15. Age estimates of duplication nodes. Histograms of age estimates for (a) the duplications mapped to the legume crown node in the Notung analysis and for duplication nodes in gene trees with only (b) Detarioideae, (c) Caesalpinioideae and (d) Papilionoideae included.

Appendix V Supplementary Information for Chapter III

Table S1. Voucher details, repository accession numbers and sequencing results for the 122 accessions used in this study.

Figure S1. ML tree of the concatenated amino acid alignment of the 510 gene alignments with more than half of the accessions present, inferred with the LG4X model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S2. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S3. ML phylogeny of 72 protein coding genes from the chloroplast genome inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S4. ML topology of concatenated alignment of 1,767 gene alignments, with ICA values indicated as branch labels. These values were calculated from gene trees of the same 1,767 gene alignments (except with short sequences removed per gene alignment), taking only those bipartitions that received at least 80% BS into account.

Figure S5. ML topology of the concatenated alignment of the 510 gene alignments with more than half of the accessions present, with number of concordant and conflicting gene trees from the same set of 510 alignments written above and below internodes, respectively. Pie charts show the number of concordant bipartitions in blue, the most common conflicting bipartition in green, all other conflicting bipartitions in red and non-informative gene trees in grey. Only bipartitions with at least 50% BS were taken into account.

Figure S6. ASTRAL tree with polytomy test results indicated, only showing non-zero p-values, for nodes with a p-value >0.05 (shown in red) a polytomy is not rejected. Terminal branch lengths are set at 1 (instead of 0) for better visualization.

Figure S7. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, but with *Cedrelinga cateniformis* removed, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S8. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, but with *Cedrelinga cateniformis* and the Samanea clade removed, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Table S1. Voucher details, repository accession numbers and sequencing results for the 122 accessions used in this study.

Taxon	Voucher	ENA accession Number	Total number Of reads	Reads on target	Number of targets Recovered	Number of gene Alignments
Abarema cochliacarpos	L.P. de Queiroz 15538 (HUEFS)	XXXXXXXX	23779774	19277048 (81.06%)	940 (97.51%)	1078 (56.29%)
Abarema jupunba	M.F. Simon 1600 (CEN)	XXXXXXXX	16357084	13117719 (80.20%)	945 (98.03%)	1042 (54.41%)
Acacia longifolia	E. Koenen 182 (Z)	XXXXXXXX	4142738	288693 (6.97%)	830 (86.10%)	520 (27.15%)
Acaciella villosa	C.E. Hughes 2635 (FHO)	XXXXXXXX	3393924	129207 (3.81%)	644 (66.80%)	274 (14.31%)
Adenanthera pavonina	Ambriansyah & Arifin AA295 (K)	XXXXXXXX	4812194	251953 (5.24%)	879 (91.18%)	606 (31.64%)
Adenopodia patens	Sandoval MS343 (K)	XXXXXXXX	7863140	791520 (10.07%)	919 (95.33%)	769 (40.16%)
Adenopodia scelerata	C. Jongkind 10602 (WAG)	XXXXXXXX	10094360	1357624 (13.45%)	950 (98.55%)	902 (47.10%)
Alantsilodendron pilosum	E. Koenen 203 (Z)	XXXXXXXX	3468930	364779 (10.52%)	933 (96.78%)	806 (42.09%)
Albizia adianthifolia	J.J. Wieringa 6278 (WAG)	XXXXXXXX	10128880	1151955 (11.37%)	945 (98.03%)	976 (50.97%)
Albizia anthelmintica	O. Maurin 0363 (JRAU)	XXXXXXXX	5576920	754863 (13.54%)	934 (96.89%)	894 (46.68%)
Albizia atakataka	E. Koenen 229 (Z)	XXXXXXXX	47608874	33141289 (69.61%)	940 (97.51%)	1010 (52.74%)
Albizia aurisparsa	E. Koenen 230 (Z)	XXXXXXXX	15816078	2127848 (13.45%)	952 (98.76%)	1047 (54.67%)
Albizia bernieri	E. Koenen 354 (Z)	XXXXXXXX	3752342	331728 (8.84%)	898 (93.15%)	692 (36.14%)
Albizia boivinii	E. Koenen 270 (Z)	XXXXXXXX	3677634	453433 (12.33%)	925 (95.95%)	882 (46.06%)
Albizia brevifolia	O. Maurin 0826 (JRAU)	XXXXXXXX	2819456	327264 (11.61%)	878 (91.08%)	678 (35.40%)
Albizia burkartiana	Stival-Santos 678 (RB)	XXXXXXXX	6094776	465465 (7.64%)	927 (96.16%)	882 (46.06%)
Albizia edwallii	Dalmaso 272 (RB)	XXXXXXXX	3339160	359353 (10.76%)	931 (96.58%)	947 (49.45%)
Albizia ferruginea	C. Jongkind 10762 (WAG)	XXXXXXXX	6722992	1035687 (15.41%)	936 (97.10%)	968 (50.55%)
Albizia grandibracteata	E. Koenen 159 (WAG)	XXXXXXXX	38109552	27944549 (73.33%)	946 (98.13%)	950 (49.61%)
Albizia inundata	J.R.I. Wood 26530	XXXXXXXX	35775492	25641806 (71.67%)	942 (97.72%)	965 (50.39%)
Albizia mahalao	E. Koenen 216 (Z)	XXXXXXXX	70271424	55605048 (79.13%)	946 (98.13%)	906 (47.31%)
Albizia masikororum	E. Koenen 237 (Z)	XXXXXXXX	12599676	1562117 (12.40%)	953 (98.86%)	1024 (53.47%)
Albizia obbiadensis	Thulin 4163 (UPS)	XXXXXXXX	5614760	735383 (13.10%)	940 (97.51%)	937 (48.93%)
Albizia polyphylla	E. Koenen 256 (Z)	XXXXXXXX	3215066	434008 (13.50%)	932 (96.68%)	843 (44.02%)
Albizia retusa	Hyland 2732 (L)	XXXXXXXX	11996368	1476589 (12.31%)	948 (98.34%)	1004 (52.43%)
Albizia sahafariensis	E. Koenen 405 (Z)	XXXXXXXX	12994846	1600201 (12.31%)	945 (98.03%)	1011 (52.79%)
Albizia saponaria	Jobson 1041 (BH)	XXXXXXXX	39202190	27805263 (70.93%)	944 (97.93%)	998 (52.11%)

Albizia splendens	Newman 2094 (E)	XXXXXXXX	55446888	41045031 (74.03%)	947 (98.24%)	972 (50.76%)
Albizia versicolor	O. Maurin 560 (JRAU)	XXXXXXXX	66547258	53198730 (79.94%)	945 (98.03%)	1045 (54.57%)
Albizia viridis	Du Puy M251 (K)	XXXXXXXX	7260284	870430 (11.99%)	934 (96.89%)	988 (51.59%)
Albizia zygia	J.J. Wieringa 5915 (WAG)	XXXXXXXX	8003478	793032 (9.91%)	941 (97.61%)	977 (51.02%)
Amblygonocarpus andongensis	Sokpon 1451 (WAG)	XXXXXXXX	5307456	263884 (4.97%)	843 (87.45%)	569 (29.71%)
Anadenanthera colubrina	L.P. de Queiroz 15685 (HUEFS)	XXXXXXXX	4286504	491557 (11.47%)	929 (96.37%)	841 (43.92%)
Archidendron lucidum	Wang and Lin 2534 (L)	XXXXXXXX	6285326	658012 (10.47%)	939 (97.41%)	972 (50.76%)
Archidendropsis granulosa	McKee 38353 (L)	XXXXXXXX	13150138	1492706 (11.35%)	947 (98.24%)	1047 (54.67%)
Aubrevillea kerstingii	Nimba Botanic Team JR957 (WAG)	XXXXXXXX	6327042	343767 (5.43%)	936 (97.10%)	770 (40.21%)
Balizia pedicellaris	L.P. de Queiroz 15529 (HUEFS)	XXXXXXXX	28193862	22668050 (80.40%)	941 (97.61%)	1104 (57.65%)
Balizia sp_nov	M.P. Morim 577 (RB)	XXXXXXXX	21239644	16903890 (79.59%)	936 (97.10%)	1071 (55.93%)
Blanchetiodendron blanchetii	L.P. de Queiroz 15616 (HUEFS)	XXXXXXXX	6639992	780827 (11.76%)	936 (97.10%)	965 (50.39%)
Calliandra hygrophila	L.P. de Queiroz 15542 (HUEFS)	XXXXXXXX	4127232	483827 (11.72%)	910 (94.40%)	732 (38.22%)
Calpocalyx dinklagei	J.J. Wieringa 6094 (WAG)	XXXXXXXX	11391816	614443 (5.39%)	929 (96.37%)	671 (35.04%)
Cathormion altissimum	C. Jongkind 10709 (WAG)	XXXXXXXX	23622566	18312807 (77.52%)	943 (97.82%)	1097 (57.28%)
Cathormion obliquifoliolatum	J.J. Wieringa 6519 (WAG)	XXXXXXXX	13303218	10816943 (81.31%)	941 (97.61%)	1047 (54.67%)
Cathormion umbellatum	Jobson 1037 (BH)	XXXXXXXX	26129888	20718828 (79.29%)	944 (97.93%)	1118 (58.38%)
Cedrelinga cateniformis	T.D. Pennington 17761 (K)	XXXXXXXX	4070738	406653 (9.99%)	919 (95.33%)	803 (41.93%)
Chidlowia sanguinea	J.J. Wieringa 4338 (WAG)	XXXXXXXX	9263792	438049 (4.73%)	888 (92.12%)	584 (30.50%)
Chloroleucon tenuiflorum	L.P. de Queiroz 15514 (HUEFS)	XXXXXXXX	7301118	779106 (10.67%)	945 (98.03%)	1031 (53.84%)
Cojoba arborea	M.F. Simon 1545 (CEN)	XXXXXXXX	9948972	1062718 (10.68%)	954 (98.96%)	1095 (57.18%)
Cylicodiscus gabunensis	M. Sosef 645A (WAG)	XXXXXXXX	6792968	649666 (9.56%)	951 (98.65%)	943 (49.24%)
Desmanthus leptophyllus	C.E. Hughes 2035 (FHO)	XXXXXXXX	4816620	392291 (8.14%)	923 (95.75%)	816 (42.61%)
Dichrostachys cinerea	O. Maurin 256 (JRAU)	XXXXXXXX	4876856	416124 (8.53%)	935 (96.99%)	822 (42.92%)
Dimorphandra macrostachya	J.R. Igançi 877 (RB)	XXXXXXXX	6731034	248839 (3.70%)	935 (96.99%)	689 (35.98%)
Diptychandra aurantiaca	J.R.I. Wood 26513	XXXXXXXX	8520962	117138 (1.37%)	881 (91.39%)	400 (20.89%)
Ebenopsis confinis	C.E. Hughes 1539 (FHO)	XXXXXXXX	5779758	654578 (11.33%)	936 (97.10%)	927 (48.41%)
Elephantorrhiza elephantina	KMS198 (JRAU)	XXXXXXXX	7379446	717080 (9.72%)	946 (98.13%)	765 (39.95%)
Entada rheedei	E. Koenen 496 (Z)	XXXXXXXX	8695656	531548 (6.11%)	948 (98.34%)	661 (34.52%)
Enterolobium contortisiliquum	L.P. de Queiroz 15579 (HUEFS)	XXXXXXXX	2729658	240130 (8.80%)	919 (95.33%)	868 (45.33%)

Erythrophleum ivorense	J.J. Wieringa 5487 (WAG)	XXXXXXXX	11500640	485354 (4.22%)	947 (98.24%)	719 (37.55%)
Faidherbia albida	O. Maurin 3495 (JRAU)	XXXXXXXX	6376338	734941 (11.53%)	945 (98.03%)	946 (49.40%)
Falcataria moluccana	Ambri & Arifin W826A (K)	XXXXXXXX	7669018	815087 (10.63%)	946 (98.13%)	991 (51.75%)
Fillaeopsis discophora	J.J. Wieringa 5498 (WAG)	XXXXXXXX	2259316	111269 (4.92%)	816 (84.65%)	597 (31.17%)
Havardia pallens	C.E. Hughes 2138 (FHO)	XXXXXXXX	6521266	726457 (11.14%)	943 (97.82%)	1056 (55.14%)
Hesperalbizia occidentalis	C.E. Hughes 1296 (FHO)	XXXXXXXX	5403622	788809 (14.60%)	947 (98.24%)	1032 (53.89%)
Hydrochorea corymbosa 1	F. Bonadeu 655 (RB)	XXXXXXXX	39645090	27356455 (69.00%)	943 (97.82%)	1028 (53.68%)
Hydrochorea corymbosa 2	J.R. Igançi 862 (RB)	XXXXXXXX	19909090	15987983 (80.30%)	944 (97.93%)	1071 (55.93%)
Inga alba	P.D. Coley & T.A. Kursar TAKPDC1677 (UT)	ERR776844	1658880	1363817 (82.21%)	942 (97.72%)	1062 (55.46%)
Inga edulis	P.D. Coley & T.A. Kursar TAKPDC1719 (UT)	ERR776838	1617410	1324567 (81.89%)	934 (96.89%)	1076 (56.19%)
Inga huberi	P.D. Coley & T.A. Kursar TAKPDC1755 (UT)	ERR776810	1555208	1291086 (83.02%)	937 (97.20%)	1085 (56.66%)
Inga laurina	K.G. Dexter 398 (E)	ERR776816	1612110	1374610 (85.27%)	944 (97.93%)	1053 (54.99%)
Inga stipularis	P.D. Coley & T.A. Kursar TAKPDC1856 (UT)	ERR776821	1692290	1393432 (82.34%)	940 (97.51%)	1055 (55.09%)
Inga tenuistipula	K.G. Dexter 110 (E)	ERR776831	1388002	1125394 (81.08%)	938 (97.30%)	1077 (56.24%)
Kanaloa kahoolawensis	Lorence 7380 (PTBG)	XXXXXXXX	12222002	1915460 (15.67%)	956 (99.17%)	933 (48.72%)
Lachesiodendron viridiflorum	L.P. de Queiroz 15614 (HUEFS)	XXXXXXXX	18632852	2381616 (12.78%)	957 (99.27%)	973 (50.81%)
Lemurodendron capuronii	E. Koenen 435 (Z)	XXXXXXXX	7108042	881933 (12.41%)	947 (98.24%)	1000 (52.22%)
Leucochloron bolivianum	C.E. Hughes 2608 (FHO)	XXXXXXXX	7946434	1218355 (15.33%)	950 (98.55%)	1046 (54.62%)
Leucochloron limae	MWC8250 (K)	XXXXXXXX	7767490	965594 (12.43%)	949 (98.44%)	1078 (56.29%)
Lysiloma candidum	B. Marazzi 300	XXXXXXXX	2030974	102461 (5.04%)	753 (78.11%)	428 (22.35%)
Macrosamanea amplissima	Bonadeu 663 (RB)	XXXXXXXX	2360238	217690 (9.22%)	920 (95.44%)	824 (43.03%)
Mariosousa sericea	MWC18949 (K)	XXXXXXXX	8160316	1450135 (17.77%)	951 (98.65%)	1011 (52.79%)
Mimosa grandidieri	E. Koenen 207 (Z)	XXXXXXXX	7792272	717042 (9.20%)	951 (98.65%)	795 (41.51%)
Mimosa tenuiflora	L.P. de Queiroz 15498 (HUEFS)	XXXXXXXX	6210710	475738 (7.66%)	944 (97.93%)	799 (41.72%)
Mimozyganthus carinatus	C.E. Hughes 2476 (FHO)	XXXXXXXX	8148502	817441 (10.03%)	943 (97.82%)	944 (49.30%)
Neptunia oleracea	E. Koenen 283 (Z)	XXXXXXXX	10836680	1176757 (10.86%)	945 (98.03%)	861 (44.96%)
Newtonia hildebrandtii	O. Maurin 2457 (JRAU)	XXXXXXXX	8663120	826146 (9.54%)	948 (98.34%)	914 (47.73%)
Pachyelasma tessmannii	J.J. Wieringa 5229 (WAG)	XXXXXXXX	11845384	793886 (6.70%)	954 (98.96%)	766 (40.00%)
Parapiptadenia zehntneri	L.P. de Queiroz 15692 (HUEFS)	XXXXXXXX	4446508	378119 (8.50%)	932 (96.68%)	939 (49.03%)
Pararchidendron pruinosum	Jobson 1039 (BH)	XXXXXXXX	7647352	738506 (9.66%)	952 (98.76%)	1052 (54.93%)

Paraserianthes lophantha	M. van Slageren & R. Newton MSRN648 (K)	XXXXXXXX	6378910	751573 (11.78%)	950 (98.55%)	1048 (54.73%)
Parkia panurensis	J.R. Igançi 842 (RB)	XXXXXXXX	2640302	231835 (8.78%)	907 (94.09%)	814 (42.51%)
Peltophorum africanum	E. Koenen 601 (Z)	XXXXXXXX	2910944	145733 (5.01%)	717 (74.38%)	367 (19.16%)
Pentaclethra macrophylla	Galeuchet & Balthazar 10 (Z)	XXXXXXXX	18158278	776900 (4.28%)	949 (98.44%)	734 (38.33%)
Piptadenia robusta	M. Luckow 4633 (BH)	XXXXXXXX	3486554	371485 (10.65%)	938 (97.30%)	898 (46.89%)
Piptadeniastrum africanum	E. Koenen 152 (WAG)	XXXXXXXX	8894316	514787 (5.79%)	948 (98.34%)	741 (38.69%)
Piptadeniopsis lomentifera	M. Luckow 4505 (BH)	XXXXXXXX	6399676	826642 (12.92%)	947 (98.24%)	926 (48.36%)
Pithecellobium dulce	B. Marazzi 309	XXXXXXXX	6485068	881345 (13.59%)	954 (98.96%)	1061 (55.40%)
Pityrocarpa moniliformis	J.R.I. Wood 26516	XXXXXXXX	6003692	449263 (7.48%)	938 (97.30%)	951 (49.66%)
Plathymenia reticulata	L.P. de Queiroz 15688 (HUEFS)	XXXXXXXX	2477330	209417 (8.45%)	920 (95.44%)	757 (39.53%)
Prosopidastrum globosum	M. Luckow sn (BH)	XXXXXXXX	6211352	691922 (11.14%)	946 (98.13%)	924 (48.25%)
Prosopis africana	Essou 2110 (WAG)	XXXXXXXX	11459252	1374601 (12.00%)	953 (98.86%)	860 (44.91%)
Prosopis laevigata	C.E. Hughes 2058 (FHO)	XXXXXXXX	3353428	246735 (7.36%)	920 (95.44%)	843 (44.02%)
Pseudopiptadenia contorta	L.P. de Queiroz 15582 (HUEFS)	XXXXXXXX	7625306	719618 (9.44%)	949 (98.44%)	1009 (52.69%)
Pseudoprosopis gilletii	J.J. Wieringa 6021 (WAG)	XXXXXXXX	5958100	264874 (4.45%)	931 (96.58%)	700 (36.55%)
Pseudosamanea guachapele	C.E. Hughes 1198 (FHO)	XXXXXXXX	7396824	1018670 (13.77%)	944 (97.93%)	1015 (53.00%)
Samanea dinklagei	C. Jongkind 7359 (WAG)	XXXXXXXX	2074148	144269 (6.96%)	912 (94.61%)	802 (41.88%)
Samanea saman	C.E. Hughes 421 (FHO)	XXXXXXXX	3344450	515562 (15.42%)	943 (97.82%)	1027 (53.63%)
Schleinitzia novoguineensis	Chaplin 57 84	XXXXXXXX	16565732	2799712 (16.90%)	956 (99.17%)	881 (46.01%)
Senegalia ataxacantha	C. Jongkind 10603 (WAG)	XXXXXXXX	11987764	1423870 (11.88%)	951 (98.65%)	941 (49.14%)
Senegalia sakalava	E. Koenen 215 (Z)	XXXXXXXX	12102414	990240 (8.18%)	946 (98.13%)	863 (45.07%)
Serianthes nelsonii	P. Moore 1241 (L)	XXXXXXXX	7283252	597846 (8.21%)	943 (97.82%)	1038 (54.20%)
Sphinga acatlensis	C.E. Hughes 2112 (FHO)	XXXXXXXX	9238996	1054313 (11.41%)	950 (98.55%)	1035 (54.05%)
Stryphnodendron pulcherrimum	L.P. de Queiroz 15482 (HUEFS)	XXXXXXXX	11852118	1440760 (12.16%)	954 (98.96%)	1007 (52.58%)
Tachigali odoratissima	M.P. Morim 562 (RB)	XXXXXXXX	7900532	193104 (2.44%)	925 (95.95%)	622 (32.48%)
Tetrapleura tetraptera	E. Koenen 155 (WAG)	XXXXXXXX	4276206	310727 (7.27%)	933 (96.78%)	707 (36.92%)
Vachellia tortilis	E. Koenen 603 (Z)	XXXXXXXX	5519408	614954 (11.14%)	930 (96.47%)	830 (43.34%)
Vachellia viguieri	E. Koenen 199 (Z)	XXXXXXXX	4782514	572917 (11.98%)	945 (98.03%)	902 (47.10%)
Viguieranthus glaber	E. Koenen 325 (Z)	XXXXXXXX	10569806	1179380 (11.16%)	950 (98.55%)	1004 (52.43%)
Xylia hoffmannii	E. Koenen 402 (Z)	XXXXXXXX	7511352	464326 (6.18%)	944 (97.93%)	713 (37.23%)

Zapoteca caracasana	C.E. Hughes 3071 (FHO)	XXXXXXXX	5236294	475921 (9.09%)	912 (94.61%)	608 (31.75%)
Zygia claviflora	J.R. Igançi 841 (RB)	XXXXXXXX	2422758	195154 (8.06%)	910 (94.40%)	784 (40.94%)
Zygia inaequalis	J.R. Igançi 832 (RB)	XXXXXXXX	5622222	494386 (8.79%)	931 (96.58%)	899 (46.95%)
Zygia racemosa	M.F. Simon 1658 (CEN)	XXXXXXXX	10769766	850638 (7.90%)	946 (98.13%)	1012 (52.85%)
Zygia sp mediana	P.D. Coley & T.A. Kursar Tip917 (UT)	ERR776824	1360502	1069298 (78.60%)	938 (97.30%)	1091 (56.97%)

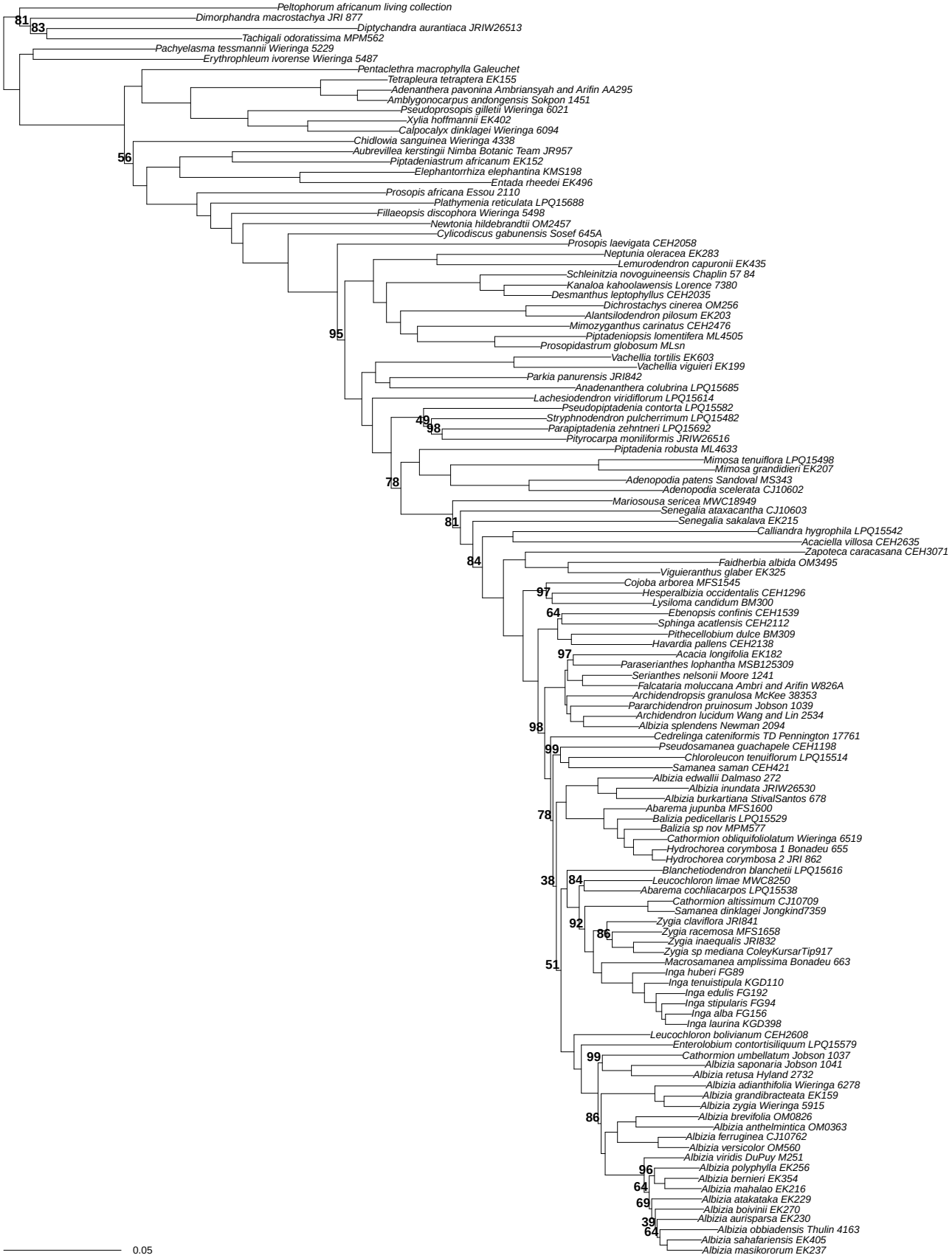


Figure S1. ML tree of the concatenated amino acid alignment of the 510 gene alignments with more than half of the accessions present, inferred with the LG4X model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S2. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S3. ML phylogeny of 72 protein coding genes from the chloroplast genome inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

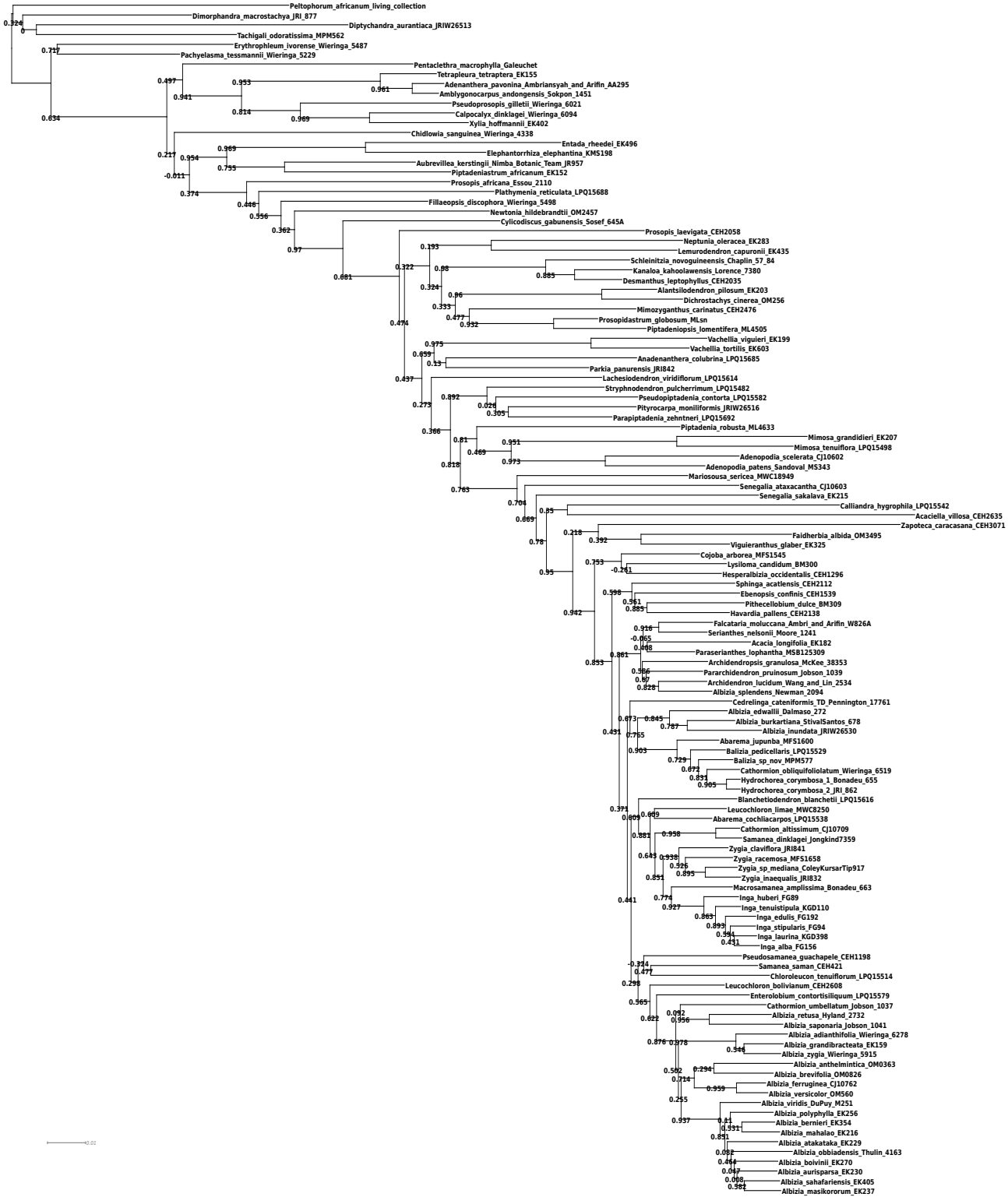


Figure S4. ML topology of concatenated alignment of 1,767 gene alignments, with ICA values indicated as branch labels. These values were calculated from gene trees of the same 1,767 gene alignments (except with short sequences removed per gene alignment), taking only those bipartitions that received at least 80% BS into account.

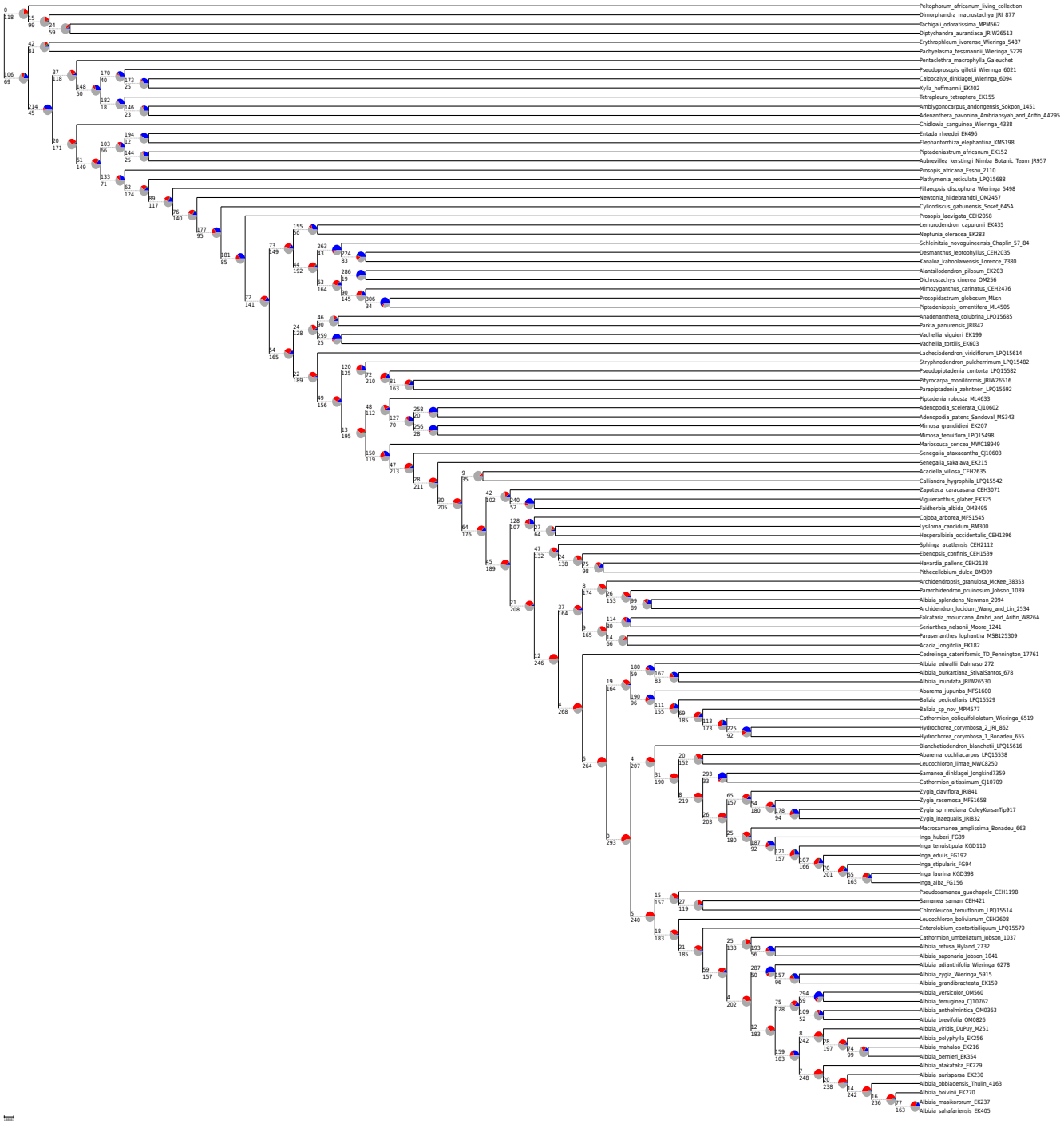


Figure S5. ML topology of the concatenated alignment of the 510 gene alignments with more than half of the accessions present, with number of concordant and conflicting gene trees from the same set of 510 alignments written above and below internodes, respectively. Pie charts show the number of concordant bipartitions in blue, the most common conflicting bipartition in green, all other conflicting bipartitions in red and non-informative gene trees in grey. Only bipartitions with at least 50% BS were taken into account.

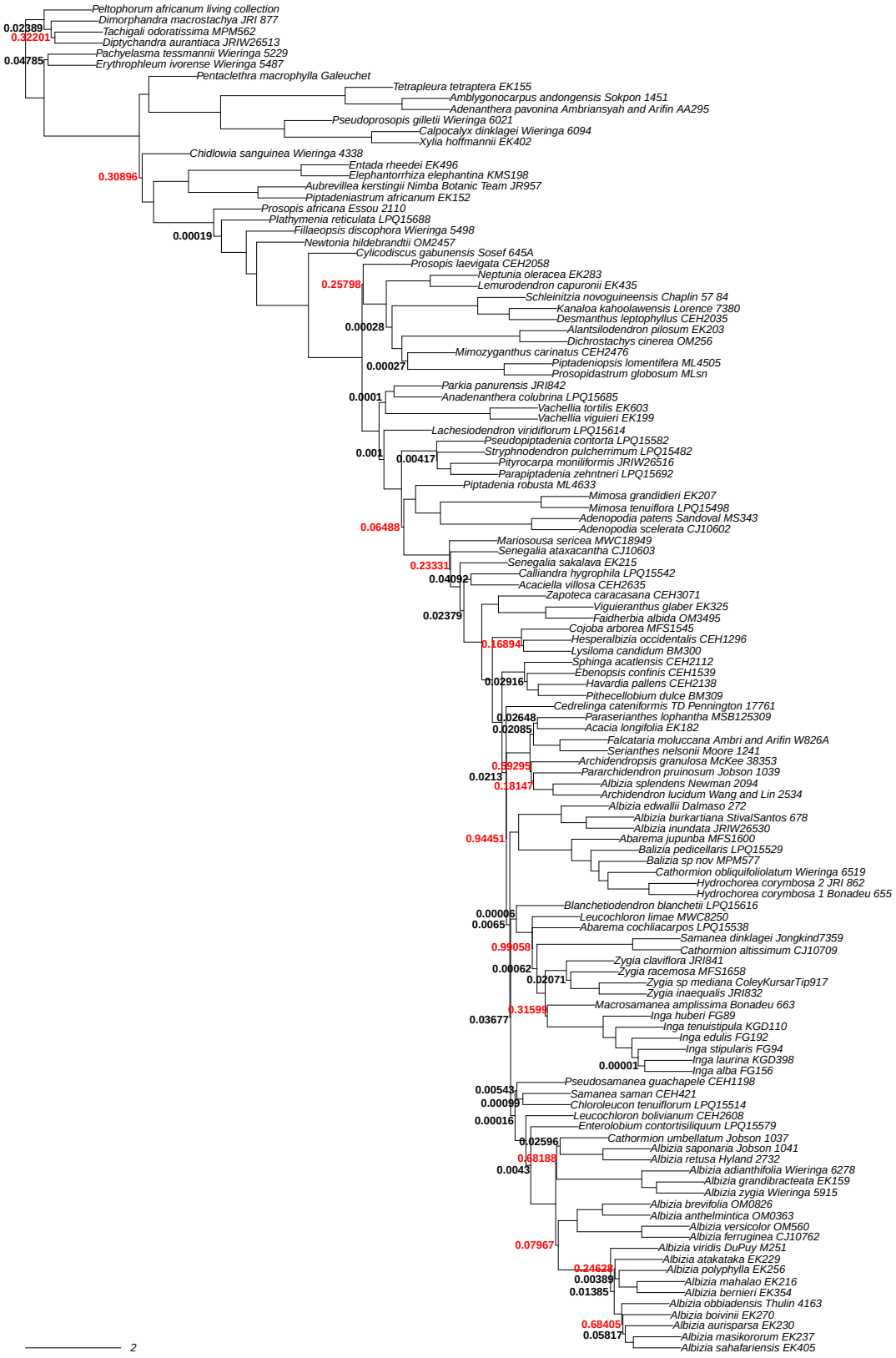


Figure S6. ASTRAL tree with polytomy test results indicated, only showing non-zero p-values, for nodes with a p-value >0.05 (shown in red) a polytomy is not rejected. Terminal branch lengths are set at 1 (instead of 0) for better visualization.

Figure S7. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, but with *Cedrelinga cateniformis* removed, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.

Figure S8. ML tree of the concatenated nucleotide alignment of the 510 gene alignments with more than half of the accessions present, but with *Cedrelinga cateniformis* and the Samanea clade removed, inferred with the GTRCAT model. Labels next to internodes indicate the bootstrap support of the subtending internode, only BS values less than 100% are shown.